Performance Evaluation in Database Research: Principles and Experiences

Ioana Manolescu\textsuperscript{1}  Stefan Manegold\textsuperscript{2}

\textsuperscript{1}INRIA Saclay–Île-de-France, France, Ioana.Manolescu@inria.fr
\textsuperscript{2}CWI Amsterdam, The Netherlands, Stefan.Manegold@cwi.nl

ICDE 2008
We are grateful to:

- EXPDB 2006 PC and participants
- VLDB 2007 Performance Evaluation participants and audience
- Dennis Shasha, SIGMOD 2008 PC chair
- SIGMOD 2008 authors of 298 (out of 436) papers who provided code for repeatability testing
- SIGMOD 2008 Repeatability Assessment committee
- Many members of the community for wide-ranging opinions and suggestions
- The ICDE 2008 organizers, in particular Malu Castellanos, Mike Carey & Qiong Luo, for providing the opportunity to present this seminar at ICDE 2008.
## Disclaimer

- There is **no single way** how to do it **right**.
- There are **many ways** how to do it **wrong**.
- This is not a “mandatory” script.
- This is more a collection of **anecdotes** or **fairy tales** — not always to be taken literally, only, but all provide some **general rules** or **guidelines** **what (not) to do**.
Planning & conducting experiments

What do you plan to do / analyze / test / prove / show?

Which data / data sets should be used?

Which workload / queries should be run?

Which hardware & software should be used?

Metrics:

What to measure?

How to measure?

How to compare?

CSI: How to find out what is going on?
Planning & conducting experiments

What do you plan to do / analyze / test / prove / show?

- Which data / data sets should be used?
- Which workload / queries should be run?
- Which hardware & software should be used?

Metrics:
- What to measure?
- How to measure?

- How to compare?
- CSI: How to find out what is going on?
Data sets & workloads

- Micro-benchmarks
- Standard benchmarks
- Real-life applications
Data sets & workloads

- Micro-benchmarks
- Standard benchmarks
- Real-life applications

- No general simple rules, which to use when
- But some guidelines for the choice...
Micro-benchmarks

**Definition**

- Specialized, stand-alone piece of software
- Isolating one particular piece of a larger system
- E.g., single DB operator (select, join, aggregation, etc.)
Micro-benchmarks

Pros

- Focused on problem at hand
- Controllable workload and data characteristics
  - Data sets (synthetic & real)
  - Data size / volume (scalability)
  - Value ranges and distribution
  - Correlation
  - Queries
  - Workload size (scalability)
- Allow broad parameter range(s)
- Useful for detailed, in-depth analysis
- Low setup threshold; easy to run
### Micro-benchmarks

<table>
<thead>
<tr>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neglect larger picture</td>
</tr>
<tr>
<td>Neglect contribution of local costs to global/total costs</td>
</tr>
<tr>
<td>Neglect impact of micro-benchmark on real-life applications</td>
</tr>
<tr>
<td>Neglect embedding in context/system at large</td>
</tr>
<tr>
<td>Generalization of result difficult</td>
</tr>
<tr>
<td>Application of insights in full systems / real-life applications no obvious</td>
</tr>
<tr>
<td>Metrics not standardized</td>
</tr>
<tr>
<td>Comparison?</td>
</tr>
</tbody>
</table>
Standard benchmarks

Examples

- RDBMS, OODBMS, ORDMBS:
  TPC-{A,B,C,H,R,DS}, OO7, ...
- XML, XPath, XQuery, XUF, SQL/XML:
  MBench, XBench, XMach-1, XMark, X007, TPoX, ...
- General Computing:
  SPEC, ...
- ...
Standard benchmarks

Pros

- Mimic real-life scenarios
- Publicly available
- Well defined (in theory ...)
- Scalable data sets and workloads (if well designed ...)
- Metrics well defined (if well designed ...)
- Easily comparable (?)
Cons

- Often “outdated” (standardization takes (too?) long)
- Often compromises
- Often very large and complicated to run
- Limited dataset variation
- Limited workload variation
- Systems are often optimized for the benchmark(s), only!
Real-life applications

Pros

- There are so many of them
- Existing problems and challenges
Real-life applications

Cons

- There are so many of them
- Proprietary datasets and workloads
Two types of experiments

Analysis: “CSI”
- Investigate (all?) details
- Analyze and understand behavior and characteristics
- Find out where the time goes and why!

Publication
- “Sell your story”
- Describe picture at large
- Highlight (some) important / interesting details
- Compare to others
Choosing the hardware

Choice mainly depends on your problem, knowledge, background, taste, etc.

What ever is required by / adequate for your problem

A laptop might not be the most suitable / representative database server...
Choosing the software

Which DBMS to use?

**Commercial**
- Require license
- “Free” versions with limited functionality and/or optimization capabilities?
- Limitations on publishing results
- No access to code
- Optimizers
- Analysis & Tuning Tools

**Open source**
- Freely available
- No limitations on publishing results
- Access to source code
Choosing the software

Other choices depend on your problem, knowledge, background, taste, etc.

- Operating system
- Programming language
- Compiler
- Scripting languages
- System tools
- Visualization tools
Metrics: What to measure?

- Basic
  - Throughput: queries per time
  - Evaluation time
    - wall-clock ("real")
    - CPU ("user")
    - I/O ("system")
    - Server-side vs. client-side
  - Memory and/or storage usage / requirements

- Comparison
  - Scale-up
  - Speed-up

- Analysis
  - System events & interrupts
  - Hardware events
Metrics: What to measure?

- Laptop: 1.5 GHz Pentium M (Dothan), 2 MB L2 cache, 2 GB RAM, 5400 RPM disk
- TPC-H ($sf = 1$)
- MonetDB/SQL v5.5.0/2.23.0
- measured last of three consecutive runs

<table>
<thead>
<tr>
<th></th>
<th>server</th>
<th>client</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>user</td>
<td>real</td>
<td>real</td>
</tr>
<tr>
<td>Q</td>
<td>real</td>
<td>real</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2830</td>
<td>3533</td>
<td>3534</td>
</tr>
<tr>
<td>16</td>
<td>3575</td>
<td>618</td>
<td>707</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1468</td>
</tr>
</tbody>
</table>

... time (milliseconds)
Metrics: What to measure?

- Laptop: 1.5 GHz Pentium M (Dothan), 2 MB L2 cache, 2 GB RAM, 5400 RPM disk
- TPC-H \( (sf = 1) \)
- MonetDB/SQL v5.5.0/2.23.0
- measured last of three consecutive runs

<table>
<thead>
<tr>
<th>Q</th>
<th>server</th>
<th>client</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>user</td>
<td>real</td>
</tr>
<tr>
<td>1</td>
<td>2830</td>
<td>3533</td>
</tr>
<tr>
<td>16</td>
<td>550</td>
<td>618</td>
</tr>
</tbody>
</table>

... time (milliseconds)
**Metrics: What to measure?**

- Laptop: 1.5 GHz Pentium M (Dothan), 2 MB L2 cache, 2 GB RAM, 5400 RPM disk
- TPC-H \((sf = 1)\)
- MonetDB/SQL v5.5.0/2.23.0
- measured last of three consecutive runs

<table>
<thead>
<tr>
<th>Q</th>
<th>server user file</th>
<th>server real file</th>
<th>client real file</th>
<th>client real terminal</th>
<th>result size</th>
<th>... time (milliseconds) output went to ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2830</td>
<td>3533</td>
<td>3534</td>
<td>3575</td>
<td>1.3 KB</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>550</td>
<td>618</td>
<td>707</td>
<td>1468</td>
<td>1.2 MB</td>
<td></td>
</tr>
</tbody>
</table>
Metrics: What to measure?

- Laptop: 1.5 GHz Pentium M (Dothan), 2 MB L2 cache, 2 GB RAM, 5400 RPM disk
- TPC-H ($sf = 1$)
- MonetDB/SQL v5.5.0/2.23.0
- measured last of three consecutive runs

<table>
<thead>
<tr>
<th>Q</th>
<th>server</th>
<th>client</th>
<th>result size</th>
<th>... time (milliseconds) output went to ...</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>user file</td>
<td>real file</td>
<td>real file</td>
<td>real terminal</td>
</tr>
<tr>
<td>1</td>
<td>2830</td>
<td>3533</td>
<td>3534</td>
<td>3575</td>
</tr>
<tr>
<td>16</td>
<td>550</td>
<td>618</td>
<td>707</td>
<td>1468</td>
</tr>
</tbody>
</table>

Be aware *what* you measure!
Metrics: How to measure?

Which tools, functions and/or system calls to use for measuring time?

- **Unix**: `/usr/bin/time`, shell built-in time
  - Command line tool ⇒ works with any executable
  - Reports “real”, “user” & “sys” time (*milliseconds*)
  - Measures entire process incl. start-up
  - **Note**: output format varies!

- **Unix**: `gettimeofday()`
  - System function ⇒ requires source code
  - Reports timestamp (*microseconds*)

- **Windows**: `timeGetTime()`
  - System function ⇒ requires source code
  - Reports timestamp (*milliseconds*)
  - Resolution implementation dependent; default can be as low as 10 milliseconds
Metrics: How to measure?

Use timings provided by the tested software (DBMS)

- IBM DB2
  - db2batch

- Microsoft SQLserver
  - GUI and system variables

- PostgreSQL
  - `postgresql.conf`
    - `log_statement_stats = on`
    - `log_min_duration_statement = 0`
    - `log_duration = on`

- MonetDB/XQuery & MonetDB/SQL
  - `mclient -lxquery -t`
  - `mclient -lsql -t`
  - `(PROFILE|TRACE) select ...`
Metrics: How to measure?

```bash
mclient -lxquery -t -s'1+2'

3

Trans 11.626 msec
Shred 0.000 msec
Query 6.462 msec
Print 1.934 msec
```

```bash
mclient -lsql -t PROFILE_select_1.sql

% . # table_name
% single_value # name
% tinyint # type
% 1 # length
[ 1 ]
#times real 62, user 0, system 0, 100
Timer 0.273 msec
```
How to run experiments

“We run all experiments in warm memory.”
“hot” vs. “cold”

- Depends on what you want to show / measure / analyze
- No formal definition, but “common sense”

**Cold run**

A cold run is a run of the query right after a DBMS is started and no (benchmark-relevant) data is preloaded into the system’s main memory, neither by the DBMS, nor in filesystem caches. Such a clean state can be achieved via a system reboot or by running an application that accesses sufficient (benchmark-irrelevant) data to flush filesystem caches, main memory, and CPU caches.

**Hot run**

A hot run is a run of a query such that as much (query-relevant) data is available as close to the CPU as possible when the measured run starts. This can (e.g.) be achieved by running the query (at least) once before the actual measured run starts.

- Be aware and document what you do / choose
“hot” vs. “cold”

- Laptop: 1.5 GHz Pentium M (Dothan), 2 MB L2 cache, 2 GB RAM, 5400 RPM disk
- TPC-H ($sf = 1$)
- MonetDB/SQL v5.5.0/2.23.0
- measured last of three consecutive runs

<table>
<thead>
<tr>
<th>Q</th>
<th>cold</th>
<th>hot</th>
<th>time (milliseconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2930</td>
<td>2830</td>
<td></td>
</tr>
</tbody>
</table>
“hot” vs. “cold”

- Laptop: 1.5 GHz Pentium M (Dothan), 2 MB L2 cache, 2 GB RAM, 5400 RPM disk
- TPC-H \((sf = 1)\)
- MonetDB/SQL v5.5.0/2.23.0
- measured last of three consecutive runs

<table>
<thead>
<tr>
<th>Q</th>
<th>cold user</th>
<th>hot user</th>
<th>... time (milliseconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2930</td>
<td>2830</td>
<td></td>
</tr>
</tbody>
</table>
“hot” vs. “cold” & user vs. real time

- Laptop: 1.5 GHz Pentium M (Dothan), 2 MB L2 cache, 2 GB RAM, 5400 RPM disk
- TPC-H \( (sf = 1) \)
- MonetDB/SQL v5.5.0/2.23.0
- measured last of three consecutive runs

<table>
<thead>
<tr>
<th>Q</th>
<th>cold</th>
<th>hot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>user</td>
<td>real</td>
</tr>
<tr>
<td>1</td>
<td>2930</td>
<td>13243</td>
</tr>
</tbody>
</table>
“hot” vs. “cold” & user vs. real time

- Laptop: 1.5 GHz Pentium M (Dothan), 2 MB L2 cache, 2 GB RAM, 5400 RPM disk
- TPC-H ($sf = 1$)
- MonetDB/SQL v5.5.0/2.23.0
- measured last of three consecutive runs

<table>
<thead>
<tr>
<th>Q</th>
<th>cold user</th>
<th>real</th>
<th>hot user</th>
<th>real</th>
<th>... time (milliseconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2930</td>
<td>13243</td>
<td>2830</td>
<td>3534</td>
<td></td>
</tr>
</tbody>
</table>

Be aware what you measure!
Once upon a time at CWI ...

- Two colleagues A & B each implemented one version of an algorithm, A the “old” version and B the improved “new” version.
- They ran identical experiments on identical machines, each for his code.
- Though both agreed that B’s new code should be significantly better, results were consistently worse.
Two colleagues A & B each implemented one version of an algorithm, A the “old” version and B the improved “new” version.

They ran identical experiments on identical machines, each for his code.

Though both agreed that B’s new code should be significantly better, results were consistently worse.

They tested, profiled, analyzed, argued, wondered, fought for several days ...
Of apples and oranges

Once upon a time at CWI ...

- Two colleagues A & B each implemented one version of an algorithm, A the “old” version and B the improved “new” version.
- They ran identical experiments on identical machines, each for his code.
- Though both agreed that B’s new code should be significantly better, results were consistently worse.
- They tested, profiled, analyzed, argued, wondered, fought for several days ...
- ... and eventually found out that A had compiled with optimization enabled, while B had not ...
Of apples and oranges

DBG

configure --enable-debug --disable-optimize --enable-assert

CFLAGS = "-g [-00]"

OPT

configure --disable-debug --enable-optimize --disable-assert

CFLAGS = "
-06 -fomit-frame-pointer -finline-functions
-malign-loops=4 -malign-jumps=4 -malign-functions=4
-fexpensive-optimizations -funroll-all-loops -funroll-loops
-frerun-cse-after-loop -frerun-loop-opt -DNDEBUG"

Manolescu, Manegold (INRIA, CWI)
Of apples and oranges

TPC-H queries

relative execution time: DBG/OPT
Of apples and oranges

- Compiler optimization ⇒ up to factor 2 performance difference
- DBMS configuration and tuning ⇒ factor $x$ performance difference ($2 \leq x \leq 10$?)
  - “Self-*” still research
  - Default settings often too “conservative”
  - Do you know all systems you use/compare equally well?

Our problem-specific, hand-tuned, prototype $X$ outperforms an out-of-the-box installation of a full-fledged off-the-shelf system $Y$, in particular when omitting query parsing, translation, optimization and result printing in $X$, while including them in $Y$.

“Absolutely fair” comparisons virtually impossible

But: Be at least aware of the crucial factors and their impact, and document accurately and completely what you do.
Of apples and oranges

- Compiler optimization ⇒ up to factor 2 performance difference
- DBMS configuration and tuning ⇒ factor $x$ performance difference ($2 \leq x \leq 10$?)
  - “Self-*” still research
  - Default settings often too “conservative”
  - Do you know all systems you use/compare equally well?

Our problem-specific, hand-tuned, prototype $X$ outperforms an out-of-the-box installation of a full-fledged off-the-shelf system $Y$,
Of apples and oranges

- Compiler optimization ⇒ up to factor 2 performance difference
- DBMS configuration and tuning ⇒ factor $x$ performance difference ($2 \leq x \leq 10$?)
  - “Self-*” still research
  - Default settings often too “conservative”
  - Do you know all systems you use/compare equally well?

Our problem-specific, hand-tuned, prototype $X$ outperforms an out-of-the-box installation of a full-fledged off-the-shelf system $Y$, in particular when omitting query parsing, translation, optimization and result printing in $X$, while including them in $Y$. 
Of apples and oranges

- Compiler optimization ⇒ up to factor 2 performance difference
- DBMS configuration and tuning ⇒ factor \( x \) performance difference (\( 2 \leq x \leq 10? \))
  - “Self-*)” still research
  - Default settings often too “conservative”
  - Do you know all systems you use/compare equally well?

Our problem-specific, hand-tuned, prototype \( X \) outperforms an out-of-the-box installation of a full-fledged off-the-shelf system \( Y \), in particular when omitting query parsing, translation, optimization and result printing in \( X \), while including them in \( Y \).

- “Absolutely fair” comparisons virtually impossible
- But:
  Be at least aware of the the crucial factors and their impact, and document accurately and completely what you do.
Do you know what happens?

Simple In-Memory Scan: SELECT MAX(column) FROM table

<table>
<thead>
<tr>
<th>Year</th>
<th>System</th>
<th>CPU Type</th>
<th>CPU Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>Sun LX</td>
<td>Sparc</td>
<td>50 MHz</td>
</tr>
<tr>
<td>1996</td>
<td>Sun Ultra</td>
<td>UltraSparc</td>
<td>200 MHz</td>
</tr>
<tr>
<td>1997</td>
<td>SunUltra</td>
<td>UltraSparcII</td>
<td>296 MHz</td>
</tr>
<tr>
<td>1998</td>
<td>DEC Alpha</td>
<td>Alpha</td>
<td>500 MHz</td>
</tr>
<tr>
<td>2000</td>
<td>Origin2000</td>
<td>R12000</td>
<td>300 MHz</td>
</tr>
</tbody>
</table>
Do you know what happens?

Simple In-Memory Scan: `SELECT MAX(column) FROM table`

- No disk-I/O involved
- Up to 10x improvement in CPU clock-speed

⇒ Yet hardly any performance improvement!??

Research: Always question what you see!

Standard profiling (e.g., `gcc -gp` + `gprof`) does not reveal more (in this case)

Need to dissect CPU & memory access costs

Use hardware performance counters to analyze cache-hits, -misses & memory accesses

VTune, oprofile, perfctr, perfmon2, PAPI, PCL, etc.
Do you know what happens?

Simple In-Memory Scan: \[ \text{SELECT MAX(column) FROM table} \]
- No disk-I/O involved
- Up to 10x improvement in CPU clock-speed
⇒ Yet hardly any performance improvement!??
- Research: Always question what you see!
Do you know what happens?

Simple In-Memory Scan: SELECT MAX(column) FROM table

- No disk-I/O involved
- Up to 10x improvement in CPU clock-speed

⇒ Yet hardly any performance improvement!??

- Research: Always question what you see!

- Standard profiling (e.g., ‘gcc -gp’ + ‘gprof’) does not reveal more (in this case)
Do you know what happens?

Simple In-Memory Scan: SELECT MAX(column) FROM table

- No disk-I/O involved
- Up to 10x improvement in CPU clock-speed

⇒ Yet hardly any performance improvement!??

- Research: Always question what you see!

- Standard profiling (e.g., `gcc -gp` + `gprof`) does not reveal more (in this case)

- Need to dissect CPU & memory access costs
- Use hardware performance counters to analyze cache-hits, -misses & memory accesses
- VTune, oprofile, perfctr, perfmon2, PAPI, PCL, etc.
Simple In-Memory Scan: SELECT MAX(column) FROM table

<table>
<thead>
<tr>
<th>Year</th>
<th>System</th>
<th>CPU Type</th>
<th>CPU Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>Sun LX</td>
<td>Sparc</td>
<td>50 MHz</td>
</tr>
<tr>
<td>1996</td>
<td>Sun Ultra</td>
<td>UltraSparc</td>
<td>200 MHz</td>
</tr>
<tr>
<td>1997</td>
<td>SunUltra</td>
<td>UltraSparcII</td>
<td>296 MHz</td>
</tr>
<tr>
<td>1998</td>
<td>DEC Alpha</td>
<td>Alpha</td>
<td>500 MHz</td>
</tr>
<tr>
<td>2000</td>
<td>Origin2000</td>
<td>R12000</td>
<td>300 MHz</td>
</tr>
</tbody>
</table>

- **Memory**
- **CPU**

Find out what happens!
Find out what happens!

Use info provided by the tested software (DBMS)

- IBM DB2
  - db2expln
- Microsoft SQLserver
  - GUI and system variables
- MySQL, PostgreSQL
  - **EXPLAIN** select ...
- MonetDB/SQL
  - (**EXPLAIN**|**TRACE**) select ...
Use profiling and monitoring tools

- `'gcc -gp' + 'gprof'`
  - Reports call tree, time per function and time per line
  - Requires re-compilation and static linking
- `'valgrind --tool=callgrind' + 'kcachegrind'`
  - Reports call tree, times, instructions executed and cache misses
  - Thread-aware
  - Does not require (re-)compilation
  - Simulation-based ⇒ slows down execution up to a factor 100
- Hardware performance counters
  - to analyze cache-hits, -misses & memory accesses
  - VTune, oprofile, perfctr, perfmon2, PAPI, PCL, etc.
- System monitors
  - `ps`, `top`, `iostat`, ...
TPC-H Q1 ($sf = 1$)  (AMD AthlonMP @ 1533 GHz, 1 GB RAM)

<table>
<thead>
<tr>
<th>cum.excl.</th>
<th>calls</th>
<th>ins.</th>
<th>IPC</th>
<th>function</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.9</td>
<td>11.9</td>
<td>846M</td>
<td>6</td>
<td>ut_fold_uint_pair</td>
</tr>
<tr>
<td>20.4</td>
<td>8.5</td>
<td>0.15M</td>
<td>27K</td>
<td>ut_fold_binary</td>
</tr>
<tr>
<td>26.2</td>
<td>5.8</td>
<td>77M</td>
<td>37</td>
<td>0.85 mikepy</td>
</tr>
<tr>
<td>29.3</td>
<td>3.1</td>
<td>23M</td>
<td>64</td>
<td>0.88 Item_sum_sum::update_field</td>
</tr>
<tr>
<td>32.3</td>
<td>3.0</td>
<td>6M</td>
<td>247</td>
<td>0.83 row_search_for_mysql</td>
</tr>
<tr>
<td>35.2</td>
<td>2.9</td>
<td>17M</td>
<td>79</td>
<td>0.70 Item_sum_avg::update_field</td>
</tr>
<tr>
<td>37.8</td>
<td>2.6</td>
<td>108M</td>
<td>11</td>
<td>0.60 rec_get_bit_field_I</td>
</tr>
<tr>
<td>40.3</td>
<td>2.5</td>
<td>6M</td>
<td>213</td>
<td>0.61 row_sel_store_mysql_rec</td>
</tr>
<tr>
<td>42.7</td>
<td>2.4</td>
<td>48M</td>
<td>25</td>
<td>0.52 rec_get_nth_field</td>
</tr>
<tr>
<td>45.1</td>
<td>2.4</td>
<td>60</td>
<td>19M</td>
<td>0.69 ha_print_info</td>
</tr>
<tr>
<td>47.5</td>
<td>2.4</td>
<td>5.9M</td>
<td>195</td>
<td>1.08 end_update</td>
</tr>
<tr>
<td>49.6</td>
<td>2.1</td>
<td>11M</td>
<td>89</td>
<td>0.98 field_conv</td>
</tr>
<tr>
<td>51.6</td>
<td>2.0</td>
<td>5.9M</td>
<td>16</td>
<td>0.77 Field_float::val_real</td>
</tr>
<tr>
<td>53.4</td>
<td>1.8</td>
<td>5.9M</td>
<td>14</td>
<td>1.07 Item_field::val</td>
</tr>
<tr>
<td>54.9</td>
<td>1.5</td>
<td>42M</td>
<td>17</td>
<td>0.51 row_sel_field_store_in_mysql..</td>
</tr>
<tr>
<td>56.3</td>
<td>1.4</td>
<td>36M</td>
<td>18</td>
<td>0.76 buf_frame_align</td>
</tr>
<tr>
<td>57.6</td>
<td>1.3</td>
<td>17M</td>
<td>38</td>
<td>0.80 Item_func_mul::val</td>
</tr>
<tr>
<td>59.0</td>
<td>1.4</td>
<td>25M</td>
<td>25</td>
<td>0.62 pthread_mutex_unlock</td>
</tr>
<tr>
<td>60.2</td>
<td>1.2</td>
<td>206M</td>
<td>2</td>
<td>0.75 hash_get_nth_cell</td>
</tr>
<tr>
<td>61.4</td>
<td>1.2</td>
<td>25M</td>
<td>21</td>
<td>0.65 mutex_test_and_set</td>
</tr>
<tr>
<td>62.4</td>
<td>1.0</td>
<td>102M</td>
<td>4</td>
<td>0.62 rec_get_1byte_offs_flag</td>
</tr>
<tr>
<td>63.4</td>
<td>1.0</td>
<td>53M</td>
<td>9</td>
<td>0.58 rec_1_get_field_start_offs</td>
</tr>
<tr>
<td>64.3</td>
<td>0.9</td>
<td>42M</td>
<td>11</td>
<td>0.65 rec_get_nth_field_extem_bit</td>
</tr>
<tr>
<td>65.3</td>
<td>1.0</td>
<td>11M</td>
<td>38</td>
<td>0.80 Item_func_minus::val</td>
</tr>
<tr>
<td>65.8</td>
<td>0.5</td>
<td>5.9M</td>
<td>38</td>
<td>0.80 Item_func_plus::val</td>
</tr>
</tbody>
</table>

MySQL gprof trace

MonetDB/MIL trace

MIL statement

<table>
<thead>
<tr>
<th>SF=1</th>
<th>SF=0.001</th>
<th>tot</th>
<th>res</th>
<th>(BW = MB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ms</td>
<td>BW</td>
<td>us</td>
<td>BW</td>
<td>MB</td>
</tr>
<tr>
<td>127</td>
<td>352</td>
<td>150</td>
<td>305</td>
<td>45</td>
</tr>
<tr>
<td>134</td>
<td>505</td>
<td>113</td>
<td>608</td>
<td>68</td>
</tr>
<tr>
<td>134</td>
<td>506</td>
<td>113</td>
<td>608</td>
<td>68</td>
</tr>
<tr>
<td>235</td>
<td>483</td>
<td>129</td>
<td>887</td>
<td>114</td>
</tr>
<tr>
<td>233</td>
<td>488</td>
<td>130</td>
<td>881</td>
<td>114</td>
</tr>
<tr>
<td>232</td>
<td>489</td>
<td>127</td>
<td>901</td>
<td>114</td>
</tr>
<tr>
<td>134</td>
<td>507</td>
<td>104</td>
<td>660</td>
<td>68</td>
</tr>
<tr>
<td>290</td>
<td>155</td>
<td>324</td>
<td>141</td>
<td>45</td>
</tr>
<tr>
<td>329</td>
<td>136</td>
<td>368</td>
<td>124</td>
<td>45</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>206</td>
<td>440</td>
<td>60</td>
<td>1527</td>
<td>91</td>
</tr>
<tr>
<td>210</td>
<td>432</td>
<td>51</td>
<td>1796</td>
<td>91</td>
</tr>
<tr>
<td>274</td>
<td>498</td>
<td>83</td>
<td>1655</td>
<td>137</td>
</tr>
<tr>
<td>274</td>
<td>499</td>
<td>84</td>
<td>1653</td>
<td>137</td>
</tr>
<tr>
<td>165</td>
<td>271</td>
<td>121</td>
<td>378</td>
<td>45</td>
</tr>
<tr>
<td>165</td>
<td>271</td>
<td>125</td>
<td>366</td>
<td>45</td>
</tr>
<tr>
<td>163</td>
<td>273</td>
<td>128</td>
<td>357</td>
<td>45</td>
</tr>
<tr>
<td>163</td>
<td>273</td>
<td>128</td>
<td>357</td>
<td>45</td>
</tr>
<tr>
<td>144</td>
<td>151</td>
<td>107</td>
<td>214</td>
<td>22</td>
</tr>
<tr>
<td>112</td>
<td>196</td>
<td>145</td>
<td>157</td>
<td>22</td>
</tr>
<tr>
<td>3724</td>
<td>2327</td>
<td></td>
<td></td>
<td>TOTAL</td>
</tr>
</tbody>
</table>
The purpose

Design measurement and simulation experiments to provide the most information with the least effort

Scenario:

- 5 parameters, each has between 10 and 40 values
- What to do?
  1. Ignore 4 parameters (!)
  2. Perform $10^5$ experiments
  3. Anything better?...

Experiment design terminology

Response  measure result

Factor    any variable that affects the response variable: parameter to be set, or environment (outer) variable

Levels    of a factor: possible values

Effect    change in the response variable due to factor level change

Replication how many times the experiment was performed

Interaction two factors interact if the effect of one depends on the level of another

Design    choice of experiments, factor level combinations and replication for each experiment
Assume two factors, $A$ and $B$, with levels $\{A_1, A_2\}$ resp. $\{B_1, B_2\}$.

- **(a)**
  
  $B_1$
  $A_1 | 3$
  $A_2 | 5$
  $B_2$
  $6 | 8$

  Same effect of $A$ change regardless of $B$

- **(b)**
  
  $B_1$
  $A_1 | 3$
  $A_2 | 5$
  $B_2$
  $6 | 9$

  Different effect of $A$ change depending on $B$

- **No interaction**
  
  - Response variable
    
    $A_1$ $A_2$
    $B_1$ $B_2$

- **Interaction**
  
  - Response variable
    
    $A_1$ $A_2$
    $B_1$ $B_2$
Common mistakes

1. Variation due to experimental error is ignored: the variation due to a factor must be compared to that due of errors!
2. Important parameters are not controlled
3. Effects of different factors are not isolated (varying many factors simultaneously)
4. Simple one-at-a-time experiment design: equally meaningful results can be obtained with less (to be seen)
5. Interactions are ignored
6. Too many experiments are conducted (enormous design). Recommended: two-stage approach
   - First experiments help identify meaningful factors and levels
   - Then conduct detailed experiments
Assume $k$ factors, such that the $i$-th factor has $n_i$ levels. Fix a common configuration and vary one factor at a time. This requires $n = 1 + \sum_{i=1}^k (n_i - 1)$ experiments.
Assume $k$ factors, such that the $i$-th factor has $n_i$ levels. Fix a common configuration and vary one factor at a time. This requires $n = 1 + \sum_{i=1}^{k} (n_i - 1)$ experiments.
Assume $k$ factors, such that the $i$-th factor has $n_i$ levels. Fix a common configuration and vary one factor at a time. This requires $n = 1 + \sum_{i=1}^{k} (n_i - 1)$ experiments.

Impossible to identify interactions (when one parameter varies, the others are constant)
Classical designs: Full factorial design

Test all possible level combinations. This requires

\[ n = 1 + \prod_{i=1}^{k} (n_i) \] experiments.
Classical designs: Full factorial design

Test all possible level combinations. This requires

\[ n = 1 + \prod_{i=1}^{k} (n_i) \] experiments.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( f_1 )</td>
<td>( f_2 )</td>
<td>( f_3 )</td>
<td>...</td>
<td>( f_k )</td>
<td>( r )</td>
</tr>
<tr>
<td>( c_1^1 )</td>
<td>( c_2^1 )</td>
<td>( c_3^1 )</td>
<td>...</td>
<td>( c_k^1 )</td>
<td>( r_0 )</td>
</tr>
<tr>
<td>( ... )</td>
<td>( ... )</td>
<td>( ... )</td>
<td>...</td>
<td>( ... )</td>
<td>( ... )</td>
</tr>
<tr>
<td>( c_1^{n_1} )</td>
<td>( c_2^1 )</td>
<td>( c_3^1 )</td>
<td>...</td>
<td>( c_k^1 )</td>
<td>( r_{n_1} )</td>
</tr>
<tr>
<td>( c_1^1 )</td>
<td>( c_2^2 )</td>
<td>( c_3^1 )</td>
<td>...</td>
<td>( c_k^1 )</td>
<td>( r_{n_1+1} )</td>
</tr>
<tr>
<td>( ... )</td>
<td>( ... )</td>
<td>( ... )</td>
<td>...</td>
<td>( ... )</td>
<td>( ... )</td>
</tr>
<tr>
<td>( c_1^1 )</td>
<td>( c_2^{n_2} )</td>
<td>( c_3^1 )</td>
<td>...</td>
<td>( c_k^1 )</td>
<td>( r_{n_1+n_2} )</td>
</tr>
<tr>
<td>( c_2^2 )</td>
<td>( c_2^2 )</td>
<td>( c_3^1 )</td>
<td>...</td>
<td>( c_k^1 )</td>
<td>( r_{n_1+n_2+1} )</td>
</tr>
<tr>
<td>( ... )</td>
<td>( ... )</td>
<td>( ... )</td>
<td>...</td>
<td>( ... )</td>
<td>( ... )</td>
</tr>
<tr>
<td>( c_2^1 )</td>
<td>( c_2^{n_2} )</td>
<td>( c_3^1 )</td>
<td>...</td>
<td>( c_k^1 )</td>
<td>( r_{n_1+2\times n_2} )</td>
</tr>
<tr>
<td>( ... )</td>
<td>( ... )</td>
<td>( ... )</td>
<td>...</td>
<td>( ... )</td>
<td>( ... )</td>
</tr>
</tbody>
</table>
Classical designs: Full factorial design

Test all possible level combinations. This requires
\[ n = 1 + \prod_{i=1}^{k} (n_i) \] experiments.

<table>
<thead>
<tr>
<th>( f_1 )</th>
<th>( f_2 )</th>
<th>( f_3 )</th>
<th>( \ldots )</th>
<th>( f_k )</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_1^1 )</td>
<td>( c_2^1 )</td>
<td>( c_3^1 )</td>
<td>( \ldots )</td>
<td>( c_k^1 )</td>
<td>( r_0 )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( c_1^{n_1} )</td>
<td>( c_2^1 )</td>
<td>( c_3^1 )</td>
<td>( \ldots )</td>
<td>( c_k^1 )</td>
<td>( r_{n_1} )</td>
</tr>
<tr>
<td>( c_1^1 )</td>
<td>( c_2^2 )</td>
<td>( c_3^1 )</td>
<td>( \ldots )</td>
<td>( c_k^1 )</td>
<td>( r_{n_1+1} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( c_1^1 )</td>
<td>( c_2^{n_2} )</td>
<td>( c_3^1 )</td>
<td>( \ldots )</td>
<td>( c_k^1 )</td>
<td>( r_{n_1+n_2} )</td>
</tr>
<tr>
<td>( c_2^2 )</td>
<td>( c_2^2 )</td>
<td>( c_3^1 )</td>
<td>( \ldots )</td>
<td>( c_k^1 )</td>
<td>( r_{n_1+n_2+1} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( c_2^1 )</td>
<td>( c_2^{n_2} )</td>
<td>( c_3^1 )</td>
<td>( \ldots )</td>
<td>( c_k^1 )</td>
<td>( r_{n_1+2\times n_2} )</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
</tbody>
</table>

Too many tests (especially if some factors are independent)
Assume $k$ factors, such that the $i$-th factor has 2 levels. This requires $n = 2^k$ experiments.

Very useful for a first-cut analysis!
### Classical designs: fractional factorial designs

**Smart** selection of level combinations

<table>
<thead>
<tr>
<th>Experiment Number</th>
<th>CPU</th>
<th>Memory Level</th>
<th>Workload Type</th>
<th>Educational Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6800</td>
<td>512 K</td>
<td>Managerial</td>
<td>High school</td>
</tr>
<tr>
<td>2</td>
<td>6800</td>
<td>2 M</td>
<td>Scientific</td>
<td>Postgraduate</td>
</tr>
<tr>
<td>3</td>
<td>6800</td>
<td>8 M</td>
<td>Secretarial</td>
<td>College</td>
</tr>
<tr>
<td>1</td>
<td>Z80</td>
<td>512 K</td>
<td>Scientific</td>
<td>College</td>
</tr>
<tr>
<td>2</td>
<td>Z80</td>
<td>2 M</td>
<td>Secretarial</td>
<td>High school</td>
</tr>
<tr>
<td>3</td>
<td>Z80</td>
<td>8 M</td>
<td>Managerial</td>
<td>Postgraduate</td>
</tr>
<tr>
<td>1</td>
<td>8086</td>
<td>512 K</td>
<td>Secretarial</td>
<td>Postgraduate</td>
</tr>
<tr>
<td>2</td>
<td>8086</td>
<td>2 M</td>
<td>Managerial</td>
<td>College</td>
</tr>
<tr>
<td>3</td>
<td>8086</td>
<td>8 M</td>
<td>Scientific</td>
<td>High school</td>
</tr>
</tbody>
</table>

Less experiments

Some information loss (interactions!) Maybe they were negligible?

Manolescu, Manegold (INRIA, CWI)
Classical designs: fractional factorial designs

Smart selection of level combinations

<table>
<thead>
<tr>
<th>Experiment Number</th>
<th>CPU</th>
<th>Memory Level</th>
<th>Workload Type</th>
<th>Educational Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6800</td>
<td>512 K</td>
<td>Managerial</td>
<td>High school</td>
</tr>
<tr>
<td>2</td>
<td>6800</td>
<td>2 M</td>
<td>Scientific</td>
<td>Postgraduate</td>
</tr>
<tr>
<td>3</td>
<td>6800</td>
<td>8 M</td>
<td>Secretarial</td>
<td>College</td>
</tr>
<tr>
<td>1</td>
<td>Z80</td>
<td>512 K</td>
<td>Scientific</td>
<td>College</td>
</tr>
<tr>
<td>2</td>
<td>Z80</td>
<td>2 M</td>
<td>Secretarial</td>
<td>High school</td>
</tr>
<tr>
<td>3</td>
<td>Z80</td>
<td>8 M</td>
<td>Managerial</td>
<td>Postgraduate</td>
</tr>
<tr>
<td>1</td>
<td>8086</td>
<td>512 K</td>
<td>Secretarial</td>
<td>Postgraduate</td>
</tr>
<tr>
<td>2</td>
<td>8086</td>
<td>2 M</td>
<td>Managerial</td>
<td>College</td>
</tr>
<tr>
<td>3</td>
<td>8086</td>
<td>8 M</td>
<td>Scientific</td>
<td>High school</td>
</tr>
</tbody>
</table>

Less experiments
Classical designs: fractional factorial designs

**Smart** selection of level combinations

<table>
<thead>
<tr>
<th>Experiment Number</th>
<th>CPU</th>
<th>Memory Level</th>
<th>Workload Type</th>
<th>Educational Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6800</td>
<td>512 K</td>
<td>Managerial</td>
<td>High school</td>
</tr>
<tr>
<td>2</td>
<td>6800</td>
<td>2 M</td>
<td>Scientific</td>
<td>Postgraduate</td>
</tr>
<tr>
<td>3</td>
<td>6800</td>
<td>8 M</td>
<td>Secretarial</td>
<td>College</td>
</tr>
<tr>
<td>1</td>
<td>Z80</td>
<td>512 K</td>
<td>Scientific</td>
<td>College</td>
</tr>
<tr>
<td>2</td>
<td>Z80</td>
<td>2 M</td>
<td>Secretarial</td>
<td>High school</td>
</tr>
<tr>
<td>3</td>
<td>Z80</td>
<td>8 M</td>
<td>Managerial</td>
<td>Postgraduate</td>
</tr>
<tr>
<td>1</td>
<td>8086</td>
<td>512 K</td>
<td>Secretarial</td>
<td>Postgraduate</td>
</tr>
<tr>
<td>2</td>
<td>8086</td>
<td>2 M</td>
<td>Managerial</td>
<td>College</td>
</tr>
<tr>
<td>3</td>
<td>8086</td>
<td>8 M</td>
<td>Scientific</td>
<td>High school</td>
</tr>
</tbody>
</table>

Less experiments
Some information loss (interactions!) Maybe they were negligible?
Example: impact of memory size and cache size on a workstation performance.

<table>
<thead>
<tr>
<th>Cache size</th>
<th>Memory size 4MB</th>
<th>Memory size 16 MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 KB</td>
<td>15</td>
<td>45</td>
</tr>
<tr>
<td>2 KB</td>
<td>25</td>
<td>75</td>
</tr>
</tbody>
</table>

Define the following variables:

\[ x_A = \begin{cases} 
-1 & \text{if 4MB memory} \\
1 & \text{if 16MB memory} 
\end{cases} \quad \text{and} \quad x_B = \begin{cases} 
-1 & \text{if 1KB cache} \\
1 & \text{if 2 KB cache} 
\end{cases} \]
Example: impact of memory size and cache size on a workstation performance.

<table>
<thead>
<tr>
<th>Cache size</th>
<th>Memory size 4MB</th>
<th>Memory size 16 MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 KB</td>
<td>15</td>
<td>45</td>
</tr>
<tr>
<td>2 KB</td>
<td>25</td>
<td>75</td>
</tr>
</tbody>
</table>

Define the following variables:

\[ x_A = \begin{cases} 
-1 & \text{if 4MB memory} \\
1 & \text{if 16MB memory} 
\end{cases} \quad \text{and} \quad x_B = \begin{cases} 
-1 & \text{if 1KB cache} \\
1 & \text{if 2 KB cache} 
\end{cases} \]

Nonlinear regression model

\[ y = q_0 + q_A x_A + q_b x_B + q_{AB} x_A x_B \]
Nonlinear regression model

\[ y = q_0 + q_A x_A + q_B x_B + q_{AB} x_A x_B \]

15 = \( q_0 - q_A - q_B + q_{AB} \)  
45 = \( q_0 + q_A - q_B - q_{AB} \)  
25 = \( q_0 - q_A + q_B - q_{AB} \)  
75 = \( q_0 + q_A + q_B + q_{AB} \)

Solving this leads to:

\[ y = 40 + 20x_A + 10x_B + 5x_A x_B \]

interpreted as: the mean is 40; the effect of memory is 20 MIPS; the effect of cache is 10 MIPS; the interaction between memory and cache accounts for 5 MIPS.
Analysis of $2^2$ factorial design

More generally:

\[y_1 = q_0 - q_A - q_B + q_{AB}\]
\[y_2 = q_0 + q_A - q_B - q_{AB}\]
\[y_3 = q_0 - q_A + q_B - q_{AB}\]
\[y_4 = q_0 + q_A + q_B + q_{AB}\]

Resolution leads to:

\[q_0 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)\]
\[q_A = \frac{1}{4}(-y_1 + y_2 - y_3 + y_4)\]
\[q_B = \frac{1}{4}(-y_1 - y_2 + y_3 + y_4)\]
\[q_{AB} = \frac{1}{4}(y_1 - y_2 - y_3 + y_4)\]
Analysis of $2^2$ factorial design

<table>
<thead>
<tr>
<th>Experiment</th>
<th>$A$</th>
<th>$B$</th>
<th>$AB$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>$y_1$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>$y_2$</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>$y_3$</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$y_4$</td>
</tr>
</tbody>
</table>

$$q_0 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)$$

$$q_A = \frac{1}{4}(-y_1 + y_2 - y_3 + y_4)$$

$$q_B = \frac{1}{4}(-y_1 - y_2 + y_3 + y_4)$$

$$q_{AB} = \frac{1}{4}(y_1 - y_2 - y_3 + y_4)$$
Analysis of $2^2$ factorial design

<table>
<thead>
<tr>
<th>Experiment</th>
<th>$A$</th>
<th>$B$</th>
<th>$AB$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>$y_1$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>$y_2$</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>$y_3$</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$y_4$</td>
</tr>
</tbody>
</table>

$$
q_0 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)
$$

$$
q_A = \frac{1}{4}(-y_1 + y_2 - y_3 + y_4)
$$

$$
q_B = \frac{1}{4}(-y_1 - y_2 + y_3 + y_4)
$$

$$
q_{AB} = \frac{1}{4}(y_1 - y_2 - y_3 + y_4)
$$
Analysis of $2^2$ factorial design

<table>
<thead>
<tr>
<th>Experiment</th>
<th>$A$</th>
<th>$B$</th>
<th>$AB$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>$y_1$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>$y_2$</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>$y_3$</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$y_4$</td>
</tr>
</tbody>
</table>

$$q_0 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)$$

$$q_A = \frac{1}{4}(-y_1 + y_2 - y_3 + y_4)$$

$$q_B = \frac{1}{4}(-y_1 - y_2 + y_3 + y_4)$$

$$q_{AB} = \frac{1}{4}(y_1 - y_2 - y_3 + y_4)$$
Analysis of $2^2$ factorial design

<table>
<thead>
<tr>
<th>Experiment</th>
<th>$A$</th>
<th>$B$</th>
<th>$AB$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>$y_1$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>$y_2$</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>$y_3$</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$y_4$</td>
</tr>
</tbody>
</table>

$$q_0 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)$$

$$q_A = \frac{1}{4}(-y_1 + y_2 - y_3 + y_4)$$

$$q_B = \frac{1}{4}(-y_1 - y_2 + y_3 + y_4)$$

$$q_{AB} = \frac{1}{4}(y_1 - y_2 - y_3 + y_4)$$
Sign table method of calculating effects

Simple algorithm to obtain $q_0, q_A, q_B, q_{AB}$ based on the sign matrix:

\[
\begin{array}{ccc}
I & A & B \\
\hline
1 & -1 & -1 \\
1 & 1 & -1 \\
1 & 1 & 1 \\
1 & -1 & 1 \\
\end{array}
\]

$\Rightarrow q_0 = \frac{1}{4} (I \cdot y)$

$q_A = \frac{1}{4} (A \cdot y)$

$q_B = \frac{1}{4} (B \cdot y)$

$q_{AB} = \frac{1}{4} (AB \cdot y)$

The mean response is $\bar{y} = q_0$. 
Simple algorithm to obtain $q_0$, $q_A$, $q_B$, $q_{AB}$ based on the sign matrix:

\[
\begin{array}{|c|c|c|c|c|}
\hline
I & A & B & AB & y \\
\hline
1 & -1 & -1 & 1 & y_1 \\
1 & 1 & -1 & -1 & y_2 \\
1 & -1 & 1 & -1 & y_3 \\
1 & 1 & 1 & 1 & y_4 \\
\hline
\end{array}
\]

\[ q_0 = \frac{1}{4} (I \cdot y) \]
\[ q_A = \frac{1}{4} (A \cdot y) \]
\[ q_B = \frac{1}{4} (B \cdot y) \]
\[ q_{AB} = \frac{1}{4} (AB \cdot y) \]
Sign table method of calculating effects

Simple algorithm to obtain $q_0, q_A, q_B, q_{AB}$ based on the sign matrix:

<table>
<thead>
<tr>
<th>$l$</th>
<th>$A$</th>
<th>$B$</th>
<th>$AB$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>$y_1$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>$y_2$</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>$y_3$</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>$y_4$</td>
</tr>
</tbody>
</table>

$$q_0 = \frac{1}{4} (I \cdot y) \quad q_A = \frac{1}{4} (A \cdot y) \quad q_B = \frac{1}{4} (B \cdot y) \quad q_{AB} = \frac{1}{4} (AB \cdot y)$$

The mean response is $\bar{y} = q_0$. 
The total variation of $y$ or sum of squares total is:

$$SST = \sum_{i=1}^{2^2} (y_i - \bar{y})^2$$
The total variation of $y$ or sum of squares total is:

$$SST = \sum_{i=1}^{2^2} (y_i - \bar{y})^2$$

Distribute SST among the factors. For a $2^2$ design:

$$SST = 2^2 q_A^2 + 2^2 q_B^2 + 2^2 q_{AB}^2$$
The total variation of $y$ or sum of squares total is:

$$SST = \sum_{i=1}^{2^2} (y_i - \bar{y})^2$$

Distribute SST among the factors. For a $2^2$ design:

$$SST = 2^2 q_A^2 + 2^2 q_B^2 + 2^2 q_{AB}^2$$

% of variation explained by $A$: $2^2 q_A^2 / SST \sim \text{importance of } A$
1. The total variation of $y$ or sum of squares total is:

$$SST = \sum_{i=1}^{2^2} (y_i - \bar{y})^2$$

2. Distribute SST among the factors. For a $2^2$ design:

$$SST = 2^2 q_A^2 + 2^2 q_B^2 + 2^2 q_{AB}^2$$

% of variation explained by $A$: $2^2 q_A^2 / SST \sim$ importance of $A$

% of variation explained by $B$: $2^2 q_B^2 / SST \sim$ importance of $B$
1. The total variation of $y$ or sum of squares total is:

$$SST = \sum_{i=1}^{2^2} (y_i - \bar{y})^2$$

2. Distribute SST among the factors. For a $2^2$ design:

$$SST = 2^2 q_A^2 + 2^2 q_B^2 + 2^2 q_{AB}^2$$

% of variation explained by $A$: $2^2 q_A^2 / SST \sim$ importance of $A$

% of variation explained by $B$: $2^2 q_B^2 / SST \sim$ importance of $B$

% of variation explained by the interaction of $A$ and $B$: $2^2 q_{AB}^2 / SST \sim$ importance of the interaction of $A$ and $B$
Example: allocation of variation

Memory interconnection networks: \{Omega, Crossbar\}
Example: allocation of variation

Memory interconnection networks: \{Omega, Crossbar\}
Two different address reference patterns: \{Random, Matrix\}
Example: allocation of variation

Memory interconnection networks: \{Omega, Crossbar\}
Two different address reference patterns: \{Random, Matrix\}
Response variables: average throughput $T$; 90% transit time in cycles $N$; average response time $R$. 
Example: allocation of variation

Memory interconnection networks: \{Omega, Crossbar\}

Two different address reference patterns: \{Random, Matrix\}

Response variables: average throughput \(T\); 90\% transit time in cycles \(N\); average response time \(R\).

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Factor</th>
<th>Level -1</th>
<th>Level 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>Type of network</td>
<td>Crossbar</td>
<td>Omega</td>
</tr>
<tr>
<td>(B)</td>
<td>Address pattern used</td>
<td>Random</td>
<td>Matrix</td>
</tr>
</tbody>
</table>
Example: allocation of variation

Memory interconnection networks: \{Omega, Crossbar\}
Two different address reference patterns: \{Random, Matrix\}
Response variables: average throughput \(T\); 90\% transit time in cycles \(N\); average response time \(R\).

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Factor</th>
<th>Level -1</th>
<th>Level 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>Type of network</td>
<td>Crossbar</td>
<td>Omega</td>
</tr>
<tr>
<td>(B)</td>
<td>Address pattern used</td>
<td>Random</td>
<td>Matrix</td>
</tr>
</tbody>
</table>
Example: allocation of variation

Memory interconnection networks: \{Omega, Crossbar\}
Two different address reference patterns: \{Random, Matrix\}
Response variables: average throughput $T$; 90\% transit time in cycles $N$; average response time $R$.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Factor</th>
<th>Level -1</th>
<th>Level 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>Type of network</td>
<td>Crossbar</td>
<td>Omega</td>
</tr>
<tr>
<td>$B$</td>
<td>Address pattern used</td>
<td>Random</td>
<td>Matrix</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$T$</th>
<th>$N$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>0.6041</td>
<td>3</td>
<td>1.655</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>0.4220</td>
<td>5</td>
<td>2.378</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>0.7922</td>
<td>2</td>
<td>1.262</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.4717</td>
<td>4</td>
<td>2.190</td>
</tr>
</tbody>
</table>

$\Rightarrow$
Example: allocation of variation

Memory interconnection networks: \{Omega, Crossbar\}
Two different address reference patterns: \{Random, Matrix\}
Response variables: average throughput $T$; 90% transit time in cycles $N$; average response time $R$.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Factor</th>
<th>Level -1</th>
<th>Level 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>Type of network</td>
<td>Crossbar</td>
<td>Omega</td>
</tr>
<tr>
<td>$B$</td>
<td>Address pattern used</td>
<td>Random</td>
<td>Matrix</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$T$</th>
<th>$N$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>0.6041</td>
<td>3</td>
<td>1.655</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>0.4220</td>
<td>5</td>
<td>2.378</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>0.7922</td>
<td>2</td>
<td>1.262</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.4717</td>
<td>4</td>
<td>2.190</td>
</tr>
</tbody>
</table>

⇒

Variation explained (%)

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$T$</th>
<th>$N$</th>
<th>$R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_A$</td>
<td></td>
<td>17.2</td>
<td>20</td>
<td>10.9</td>
</tr>
<tr>
<td>$q_B$</td>
<td></td>
<td>77.0</td>
<td>80</td>
<td>87.8</td>
</tr>
<tr>
<td>$q_{AB}$</td>
<td></td>
<td>5.8</td>
<td>0</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Conclusion: the address pattern influences most. The chosen patterns are very different.
Example: allocation of variation

Memory interconnection networks: \{Omega, Crossbar\}
Two different address reference patterns: \{Random, Matrix\}
Response variables: average throughput $T$; 90\% transit time in cycles $N$; average response time $R$.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Factor</th>
<th>Level -1</th>
<th>Level 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>Type of network</td>
<td>Crossbar</td>
<td>Omega</td>
</tr>
<tr>
<td>$B$</td>
<td>Address pattern used</td>
<td>Random</td>
<td>Matrix</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$T$</th>
<th>$N$</th>
<th>$R$</th>
<th>Variation explained (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>0.6041</td>
<td>3</td>
<td>1.655</td>
<td>$q_A$ 17.2 20 10.9</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>0.4220</td>
<td>5</td>
<td>2.378</td>
<td>$q_B$ 77.0 80 87.8</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>0.7922</td>
<td>2</td>
<td>1.262</td>
<td>$q_{AB}$ 5.8 0 1.3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.4717</td>
<td>4</td>
<td>2.190</td>
<td></td>
</tr>
</tbody>
</table>

Conclusion: the address pattern influences most. The chosen patterns are very different.
From $2^2$ to $2^k$

$k$ factors, each with 2 levels. There will be:

- $k$ main effects
- $C_k^2$ two-factors interactions
- $C_k^3$ three-factors interactions
- ... 

\[
y = q_0 + q_{A_1} x_{A_1} + q_{A_2} x_{A_2} + \ldots + q_{A_k} x_{A_k} + q_{A_1 A_2} x_{A_1} x_{A_2} + q_{A_1 A_3} x_{A_1} x_{A_3} + \ldots + q_{A_{k-1} A_k} x_{A_{k-1}} x_{A_k} + \ldots + q_{A_1 A_2 \ldots A_k} x_{A_1} x_{A_2} \ldots x_{A_k}
\]

$2^k$ experiments allow to compute $q_0, q_{A_1}, \ldots, q_{A_1 A_2 \ldots A_k}$. Then the analysis proceeds as for $2^2$. 

Manolescu, Manegold (INRIA, CWI)
Performance Evaluation: Principles & Experiences ICDE 2008 93/1
Preparing a fractional factorial design

There are $k$ parameters, each with 2 levels. Instead of $2^k$, we aim to judiciously choose $2^{k-p}$ level combinations to test.

The net effect is simplifying the dependency model:

$$y = q_0 + q_{A_1}x_{A_1} + q_{A_2}x_{A_2} + \ldots + q_{A_k}x_{A_k} + q_{A_1A_2}x_{A_1}x_{A_2} + \ldots + q_{A_{k-1}A_k}x_{A_{k-1}}x_{A_k} + \ldots + q_{A_1A_2\ldots A_{k-1}}x_{A_1}x_{A_2}\ldots x_{A_{k-1}} + q_{A_1A_2\ldots A_k}x_{A_1}x_{A_2}\ldots x_{A_{k-1}}x_{A_k}$$
Preparing a fractional factorial design

There are $k$ parameters, each with 2 levels. Instead of $2^k$, we aim to judiciously choose $2^{k-p}$ level combinations to test. The net effect is *simplifying the dependency model*: 

$$y = q_0 + q_{A_2}A_2 + \ldots + q_{A_k}A_k + q_{A_1A_2}A_1A_2 + \ldots + q_{A_{k-1}A_k}A_{k-1}A_k + \ldots + q_{A_1A_2\ldots A_{k-1}}A_1A_2\ldots A_{k-1} + \ldots$$
Preparing a fractional factorial design

There are $k$ parameters, each with 2 levels. Instead of $2^k$, we aim to judiciously choose $2^{k-p}$ level combinations to test. The net effect is *simplifying the dependency model*:

$$
y = q_0 + q_{A_1}x_{A_1} + q_{A_2}x_{A_2} + \ldots + q_{A_k}x_{A_k} +
q_{A_1A_2}x_{A_1}x_{A_2} + q_{A_1A_3}x_{A_1}x_{A_3} + \ldots + q_{A_{k-1}A_k}x_{A_{k-1}}x_{A_k} +
\ldots + q_{A_1A_2\ldots A_{k-1}}x_{A_1}x_{A_2} \ldots x_{A_{k-1}} +
q_{A_1A_2\ldots A_k}x_{A_1}x_{A_2} \ldots x_{A_k}
$$

becomes:

$$y = q_0 + q_{A_2}x_{A_2} + \ldots + q_{A_k}x_{A_k} +
q_{A_1A_2}x_{A_1}x_{A_2} + \ldots + q_{A_{k-1}A_k}x_{A_{k-1}}x_{A_k} +
\ldots + q_{A_1A_2\ldots A_{k-1}}x_{A_1}x_{A_2} \ldots x_{A_{k-1}}$$
Preparing a fractional factorial design

There are \( k \) parameters, each with 2 levels. Instead of \( 2^k \), we aim to judiciously choose \( 2^{k-p} \) level combinations to test.

The net effect is simplifying the dependency model:

\[
y = q_0 + q_{A_1} x_{A_1} + q_{A_2} x_{A_2} + \ldots + q_{A_k} x_{A_k} +
\]
\[
q_{A_1 A_2} x_{A_1} x_{A_2} + q_{A_1 A_3} x_{A_1} x_{A_3} + \ldots + q_{A_{k-1} A_k} x_{A_{k-1}} x_{A_k} +
\]
\[
\ldots + q_{A_1 A_2 \ldots A_{k-1}} x_{A_1} x_{A_2} \ldots x_{A_{k-1}} +
\]
\[
q_{A_1 A_2 \ldots A_k} x_{A_1} x_{A_2} \ldots x_{A_k}
\]

becomes:
\[
y = q_0 + q_{A_2} x_{A_2} + \ldots + q_{A_k} x_{A_k} +
\]
\[
q_{A_1 A_2} x_{A_1} x_{A_2} + \ldots + q_{A_{k-1} A_k} x_{A_{k-1}} x_{A_k} +
\]
\[
\ldots + q_{A_1 A_2 \ldots A_{k-1}} x_{A_1} x_{A_2} \ldots x_{A_{k-1}}
\]

\( 2^{k-p} \) measures / equations / coefficients instead of \( 2^k \)
Preparing a fractional factorial design

There are $k$ parameters, each with 2 levels. Instead of $2^k$, we aim to judiciously choose $2^{k-p}$ level combinations to test. The net effect is *simplifying the dependency model*:

$$y = q_0 + q_{A_1}x_{A_1} + q_{A_2}x_{A_2} + \ldots + q_{A_k}x_{A_k} +$$

$$q_{A_1A_2} x_{A_1} x_{A_2} + q_{A_1A_3} x_{A_1} x_{A_3} + \ldots + q_{A_{k-1}A_k} x_{A_{k-1}} x_{A_k} +$$

$$\ldots + q_{A_1A_2\ldots A_{k-1}} x_{A_1} x_{A_2} \ldots x_{A_{k-1}} +$$

$$q_{A_1A_2\ldots A_k} x_{A_1} x_{A_2} \ldots x_{A_k}$$

becomes:

$$y = q_0 + q_{A_2}x_{A_2} + \ldots + q_{A_k}x_{A_k} +$$

$$q_{A_1A_2} x_{A_1} x_{A_2} + \ldots + q_{A_{k-1}A_k} x_{A_{k-1}} x_{A_k} +$$

$$\ldots + q_{A_1A_2\ldots A_{k-1}} x_{A_1} x_{A_2} \ldots x_{A_{k-1}}$$

$2^{k-p}$ measures / equations / coefficients instead of $2^k$

Ideally, the coefficients replaced with 0 are small
Preparing a fractional factorial design

We need a sign table of dimension $2^{k-p}$

- Each column consists of -1 and +1 and has the sum zero.
- Columns should be orthogonal.

Method:
1. Pick $k-p$ factors, build a full factorial design of size $2^{k-p}$.
2. Chose $p$ among the rightmost $2^{k-p} - k + p - 1$ columns and label them with the $p$ factors not chosen in step 1.
Preparing a fractional factorial design

We need a sign table of dimension $2^{k-p}$

- Each column consists of -1 and +1 and has the sum zero.
- Columns should be orthogonal.

Method:

1. Pick $k - p$ factors, build a full factorial design of size $k - p$.
2. Chose $p$ among the rightmost $2^{k-p} - k + p - 1$ columns and label them with the $p$ factors not chosen in step 1.
Preparing a fractional factorial design of $2^{7-4}$

We start with $k = 7$ factors: $A, B, C, D, E, F, G$.
We pick the first $k - p = 7 - 4 = 3$ factors: $A, B, C$ and build a full factorial design for these:

<table>
<thead>
<tr>
<th>Exp.</th>
<th>$A$</th>
<th>$B$</th>
<th>$C$</th>
<th>$AB$</th>
<th>$AC$</th>
<th>$BC$</th>
<th>$ABC$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Preparing a fractional factorial design of $2^{7-4}$

We change the names of the rightmost 4 columns into $D, E, F, G$:

<table>
<thead>
<tr>
<th>Exp.</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- 7 zero-sum columns: so that both levels get equally tested
- 3 orthogonal factor columns ($A, B$ and $C$): any two of these factors agree (product=1) as often as they disagree (product=-1)
- all coefficients of interactions have been erased.
Preparing a fractional factorial design of $2^{4-1}$

<table>
<thead>
<tr>
<th>Exp.</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>AB</th>
<th>AC</th>
<th>BC</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Assume 8 experiments lead to results $y_1, y_2, \ldots, y_8$. The confounded effects of $D$ and the $ABC$ interaction is:

$$y \cdot D = y \cdot A \cdot B \cdot C = -y_1 + y + 2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8.$$  

This particular fractional design is denoted $D=ABC$. 
Fractional design $D=ABC$ for $2^{4-1}$

If $D = ABC$, then also:

- $A \cdot D = A \cdot ABC = I \cdot BC = BC$; $AD = BC$
- $B \cdot D = B \cdot ABC = I \cdot AC = AC$; $BD = AC$
- $C \cdot D = C \cdot ABC = I \cdot AB = AB$; $AB = CD$
- $BC \cdot D = BC \cdot ABC = I \cdot A$; $A = BCD$
- Also: $B = ACD$, $C = ABD$, $I = ABCD$

This design confounds:

- the mean with the 4th order interaction;
- the main effects with 3rd order interactions.
If $D = ABC$, then also:

- $A \cdot D = A \cdot ABC = I \cdot BC = BC; \ AD = BC$
- $B \cdot D = B \cdot ABC = I \cdot AC = AC; \ BD = AC$
- $C \cdot D = C \cdot ABC = I \cdot AB = AB; \ AB = CD$
- $BC \cdot D = BC \cdot ABC = I \cdot A; \ A = BCD$
- Also: $B = ACD$, $C = ABD$, $I = ABCD$

This design confounds:

- the mean with the 4th order interaction;
- the main effects with 3rd order interactions.

The hope is that 3rd and 4th order interactions are small.
## Comparison of two $2^{4-1}$ designs

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>AB</th>
<th>AC</th>
<th>BC</th>
<th>ABC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Fractional design $D = ABC$  \rightarrow  Fractional design $D = AB$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>AB</th>
<th>AC</th>
<th>BC</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>AC</th>
<th>BC</th>
<th>ABC</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Comparison of two $2^{4-1}$ designs

Confoundings of $D = ABC$:
- $AD = BC$, $BD = AC$, $AB = CD$
- $A = BCD$, $B = ACD$, $C = ABD$
- $I = ABCD$

Confoundings of $D = AB$:
- $A = BD$, $B = AD$, $D = AB$
- $I = ABD$
- $AC = BCD$, $BC = ACD$, $CD = ABC$
- $C = ABCD$

It is assumed higher order interactions are less important than lower order interactions ("sparsity of effects" principle). Therefore, designs that confound higher order interactions are preferred.
Comparison of two $2^{4-1}$ designs

Confounding of $D = ABC$:
- $AD = BC$, $BD = AC$, $AB = CD$
- $A = BCD$, $B = ACD$, $C = ABD$
- $I = ABCD$

Confounding of $D = AB$:
- $A = BD$, $B = AD$, $D = AB$
- $I = ABD$
- $AC = BCD$, $BC = ACD$, $CD = ABC$
- $C = ABCD$

$D = ABC$ is preferred

It is assumed higher order interactions are less important than lower order interactions ("sparsity of effects" principle). Therefore, designs that confound higher order interactions are preferred.
Conclusion on experiment design

Design

Picking the factors, their levels, and the replication degree (number of repetitions)
Design

Picking the factors, their levels, and the replication degree (number of repetitions)

Design has a huge impact:

- You don’t know what you haven’t tested
- Ignoring important parameter leads to brittle results
Conclusion on experiment design

Design

Picking the factors, their levels, and the replication degree (number of repetitions)

Design has a huge impact:

- You don’t know what you haven’t tested
- Ignoring important parameter leads to brittle results
- Testing all possible level combinations is unfeasible
Conclusion on experiment design

Design

Picking the factors, their levels, and the replication degree (number of repetitions)

Design has a huge impact:

- You don’t know what you haven’t tested
- Ignoring important parameter leads to brittle results
- Testing all possible level combinations is unfeasible
- There exist standard, well-founded procedures for getting the same information with less effort:
  1. Run a $2^k$ (or a $2^k-p$) design
  2. Evaluate factor importance
  3. Pick important factors and possibly refine levels
We all know

A picture is worth a thousand words
We all know

A picture is worth a thousand words

Er, maybe not all pictures...
Graphical presentation of results

We all know

A picture is worth a thousand words

Er, maybe not all pictures...

(Borrowed from T.Grust’s slides at VLDB 2007 panel)
Guidelines for preparing good graphic charts

Require minimum effort from the reader
Guidelines for preparing good graphic charts

Require minimum effort from the reader

- Not the minimum effort from you
Guidelines for preparing good graphic charts

Require minimum effort from the reader

- Not the minimum effort from you
- Try to be honest: how would you like to see it?
Guidelines for preparing good graphic charts

Require minimum effort from the reader

- Not the minimum effort from you
- Try to be honest: how would you like to see it?

![Graphs showing response time vs. number of users for different options A, B, C.](image-url)
Guidelines for preparing good graphic charts

Maximize information: try to make the graph self-sufficient

- Use keywords in place of symbols to avoid a join in the reader’s brain
- Use informative axis labels: prefer “Average I/Os per query” to “Average I/Os” to “I/Os”
- Include units in the labels: prefer “CPU time (ms)” to “CPU time”
Guidelines for preparing good graphic charts

**Maximize information**: try to make the graph self-sufficient

- Use keywords in place of symbols to avoid a join in the reader’s brain
- Use informative axis labels: prefer “Average I/Os per query” to “Average I/Os” to “I/Os”
- Include units in the labels: prefer “CPU time (ms)” to “CPU time”

**Use commonly accepted practice**: present what people expect

- *Usually* axes begin at 0, the factor is plotted on x, the result on y
- *Usually* scales are linear, increase from left to right, divisions are equal
- Use exceptions as necessary
Guidelines for preparing good graphic charts

Minimize ink: present as much information as possible with as little ink as possible
Guidelines for preparing good graphic charts

Minimize ink: present as much information as possible with as little ink as possible
Prefer the chart that gives the most information out of the same data
Guidelines for preparing good graphic charts

Minimize ink: present as much information as possible with as little ink as possible
Prefer the chart that gives the most information out of the same data
Common presentation mistakes

Presenting too many alternatives on a single chart

Rules of thumb, to override with good reason:

- A line chart should be limited at 6 curves
- A column chart or bar should be limited to 10 bars
- A pie chart should be limited to 8 components
- Each cell in a histogram should have at least five data points
Presenting many result variables on a single chart
Commonly done to fit into available page count :-(
Common presentation mistakes

Presenting many result variables on a single chart
Commonly done to fit into available page count :-(

Huh?
Common presentation mistakes

Using symbols in place of text
Common presentation mistakes

Using symbols in place of text

![Graphs showing response time vs. arrival rate for different arrival rates and service rates.](Image)

- **Response time**
  - 1 job/sec
  - 3 jobs/sec
  - 2 jobs/sec

- **Arrival rate**
  - $\lambda$
  - $\mu=1$
  - $\mu=3$
  - $\mu=2$

Humans get frustrated by computing joins

Manolescu, Manegold (INRIA, CWI)

Performance Evaluation: Principles & Experiences

ICDE 2008
Common presentation mistakes

Using symbols in place of text

Human brain is a poor join processor

- \( \lambda \) Arrival rate
- \( R \) Response time
- \( \mu=1 \), \( \mu=2 \), \( \mu=3 \)
Common presentation mistakes

Using symbols in place of text

Human brain is a poor join processor
Humans get frustrated by computing joins
Common presentation mistakes

Change the graphical layout of a given curve from one figure to another
Common presentation mistakes

Change the graphical layout of a given curve from one figure to another

//tn time (different machines)

Q26(n) time (different machines)
Common presentation mistakes

Change the graphical layout of a given curve from one figure to another

What do you mean “my graphs are not legible”?
MINE is better than YOURS!
Pictorial games

MINE is better than YOURS!

![Graph showing performance comparison between MINE and YOURS.](image-url)
Pictorial games

MINE is better than YOURS!

A-ha
Recommended layout: let the useful height of the graph be 3/4th of its useful width
Pictorial games

Plot random quantities without confidence intervals

Manolescu, Manegold (INRIA, CWI)  Performance Evaluation: Principles & Experiences  ICDE 2008  140/1
Pictorial games

Plot random quantities without confidence intervals

Overlapping confidence intervals sometimes mean the two quantities are statistically indifferent
Manipulating cell size in histograms

Rule of thumb: each cell should have at least five points

Not sufficient to uniquely determine what one should do.
Rule of thumb: each cell should have at least five points
Not sufficient to uniquely determine what one should do.
**Pictorial games: gnuplot & **\LaTeX**

**Rule of thumb for papers:**
- width of plot = \textwidth
  - \texttt{set size ratio 0 \xtextwidth*0.75, \ytextwidth}

---

**Manolescu, Manegold (INRIA, CWI)**
**Performance Evaluation: Principles & Experiences**
**ICDE 2008**
Pictorial games: gnuplot & \LaTeX

default:
set size ratio 0 1,1

better:
set size ratio 0 0.5,0.5

Rule of thumb for papers:
width of plot = \textwidth
⇒ set size ratio 0 x *0.75, y

Manolescu, Manegold (INRIA, CWI)
Performance Evaluation: Principles & Experiences ICDE 2008 145/1
Pictorial games: gnuplot & \LaTeX

default:
\texttt{set size ratio 0 1,1}

better:
\texttt{set size ratio 0 0.5,0.5}

Rule of thumb for papers:

width of plot = $x\backslash\textwidth$
\[\Rightarrow \text{set size ratio } 0 \times 0.75, y\]
"We use a machine with 3.4 GHz."
“We use a machine with 3.4 GHz.”

⇒ Under-specified!
Specifying hardware environments

```
cat /proc/cpuinfo

processor : 0
vendor_id : GenuineIntel
cpu family : 6
model : 13
model name : Intel(R) Pentium(R) M processor 1.50GHz
stepping : 6
cpu MHz : 600.000
cache size : 2048 KB
fdiv_bug : no
hlt_bug : no
f00f_bug : no
coma_bug : no
fpu : yes
fpu_exception : yes
cpuid level : 2
wp : yes
flags : fpu vme de pse tsc msr mce cx8 mtrr pge mca cmov pat clflush
dts acpi mmx fxsr sse sse2 ss tm pbe up bts est tm2
bogomips : 1196.56
clflush size : 64
```
Specifying hardware environments

/sbin/lspci -v

00:00.0 Host bridge: Intel Corporation 82852/82855 GM/GME/PM/GMV Processor to I/O Controller (rev 02)
  Flags: bus master, fast devsel, latency 0
  Memory at <unassigned> (32-bit, prefetchable)
  Capabilities: <access denied>
  Kernel driver in use: agpgart-intel

...

01:08.0 Ethernet controller: Intel Corporation 82801DB PRO/100 VE (MOB) Ethernet Controller (rev 83)
  Subsystem: Benq Corporation Unknown device 5002
  Flags: bus master, medium devsel, latency 64, IRQ 10
  Memory at e0000000 (32-bit, non-prefetchable) [size=4K]
  I/O ports at c000 [size=64]
  Capabilities: <access denied>
  Kernel driver in use: e100
  Kernel modules: e100

/sbin/lspci -v | wc

  151 lines
  861 words
  6663 characters
Specifying hardware environments

/sbin/lspci -v

00:00.0 Host bridge: Intel Corporation 82852/82855 GM/GME/PM/GMV Processor to I/O Controller (rev 02)
   Flags: bus master, fast devsel, latency 0
   Memory at <unassigned> (32-bit, prefetchable)
   Capabilities: <access denied>
   Kernel driver in use: agpgart-intel

...

01:08.0 Ethernet controller: Intel Corporation 82801DB PRO/100 VE (MOB) Ethernet Controller (rev 83)
   Subsystem: Benq Corporation Unknown device 5002
   Flags: bus master, medium devsel, latency 64, IRQ 10
   Memory at e0000000 (32-bit, non-prefetchable) [size=4K]
   I/O ports at c000 [size=64]
   Capabilities: <access denied>
   Kernel driver in use: e100
   Kernel modules: e100

/sbin/lspci -v | wc

  151 lines
  861 words
  6663 characters

⇒ Over-specified!
Specifying hardware environments

- **CPU**: Vendor, model, generation, clockspeed, cache size(s):
  1.5 GHz Pentium M (Dothan), 32 KB L1 cache, 2 MB L2 cache

- **Main memory**: size
  2 GB RAM

- **Disk (system)**: size & speed
  120 GB Laptop ATA disk @ 5400 RPM
  1 TB striped RAID-0 system (5× 200 GB S-ATA disk @ 7200 RPM)
Specifying software environments

Product names, **exact version numbers**, and/or sources where obtained from
Making experiments repeatable

Purpose: another human equipped with the appropriate software and hardware can repeat your experiments.

- Your supervisor / your students
Making experiments repeatable

Purpose: another human equipped with the appropriate software and hardware can repeat your experiments.

- Your supervisor / your students
- Your colleagues
Making experiments repeatable

Purpose: another human equipped with the appropriate software and hardware can repeat your experiments.

- Your supervisor / your students
- Your colleagues
- Yourself, 3 months later when you have a new idea
- Yourself, 3 years later when writing the thesis or answering requests for that journal version of your conference paper
- Future researchers (you get cited!)
Making experiments repeatable

Purpose: another human equipped with the appropriate software and hardware can repeat your experiments.

- Your supervisor / your students
- Your colleagues
- Yourself, 3 months later when you have a new idea
- Yourself, 3 years later when writing the thesis or answering requests for that journal version of your conference paper
Making experiments repeatable

Purpose: another human equipped with the appropriate software and hardware can repeat your experiments.

- Your supervisor / your students
- Your colleagues
- Yourself, 3 months later when you have a new idea
- Yourself, 3 years later when writing the thesis or answering requests for that journal version of your conference paper
- Future researchers (you get cited!)
Making experiments repeatable

Purpose: another human equipped with the appropriate software and hardware can repeat your experiments.

- Your supervisor / your students
- Your colleagues
- Yourself, 3 months later when you have a new idea
- Yourself, 3 years later when writing the thesis or answering requests for that journal version of your conference paper
- Future researchers (you get cited!)

Making experiments repeatable means:

1. Making experiments **portable** and **parameterizable**
2. Building a **test suite** and scripts
3. Writing **instructions**
Making experiments portable

Try to use not-so-exotic hardware
Try to use free or commonly available tools (databases, compilers, plotters...)

Clearly, scientific needs go first (joins on graphic cards; smart card research; energy consumption study...)

You may omit using Matlab as the driving platform for the experiments
20-years old software that only works on an old SUN and is now unavailable

If you really love your code, you may even maintain it

Code maintenance
Making experiments portable

Try to use not-so-exotic hardware
Try to use free or commonly available tools (databases, compilers, plotters…)
Clearly, scientific needs go first (joins on graphic cards; smart card research; energy consumption study…)

Manolescu, Manegold (INRIA, CWI)
Performance Evaluation: Principles & Experiences
ICDE 2008
Making experiments portable

Try to use not-so-exotic hardware
Try to use free or commonly available tools (databases, compilers, plotters...)
Clearly, scientific needs go first (joins on graphic cards; smart card research; energy consumption study...)

You may omit using
Matlab as the driving platform for the experiments
Making experiments portable

Try to use not-so-exotic hardware
Try to use free or commonly available tools (databases, compilers, plotters...)
Clearly, scientific needs go first (joins on graphic cards; smart card research; energy consumption study...)

You may omit using
Matlab as the driving platform for the experiments
20-years old software that only works on an old SUN and is now unavailable
Making experiments portable

Try to use not-so-exotic hardware
Try to use free or commonly available tools (databases, compilers, plotters...)
Clearly, scientific needs go first (joins on graphic cards; smart card research; energy consumption study...)

You may omit using

Matlab as the driving platform for the experiments
20-years old software that only works on an old SUN and is now unavailable

- If you really love your code, you may even maintain it
Making experiments portable

Try to use not-so-exotic hardware
Try to use free or commonly available tools (databases, compilers, plotters...)
Clearly, scientific needs go first (joins on graphic cards; smart card research...)
You may omit using Matlab as the driving platform for the experiments
20-years old software that only works on an old SUN and is now unavailable
If you really love your code, you may even maintain it

Code maintenance
Making experiments portable

Try to use not-so-exotic hardware
Try to use free or commonly available tools (databases, compilers, plotters...)
Clearly, scientific needs go first (joins on graphic cards; smart card research; energy consumption study...)

You may omit using

Matlab as the driving platform for the experiments
20-years old software that only works on an old SUN and is now unavailable (if you really love your code, you may even maintain it)
4-years old library that is no longer distributed and you do no longer have (idem)
Try to use not-so-exotic hardware
Try to use free or commonly available tools (databases, compilers, plotters...)
Clearly, scientific needs go first (joins on graphic cards; smart card research; energy consumption study...)

You may omit using
Matlab as the driving platform for the experiments
20-years old software that only works on an old SUN and is now unavailable (if you really love your code, you may even maintain it)
4-years old library that is no longer distributed and you do no longer have (idem)
/usr/bin/time to time execution, parse the output with perl, divide by zero
Which abstract do you prefer?

Abstract (Take 1)
We provide a new algorithm that consistently outperforms the state of the art.
<table>
<thead>
<tr>
<th>Abstract (Take 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>We provide a new algorithm that consistently outperforms the state of the art.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Abstract (Take 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>We provide a new algorithm that on a Debian Linux machine with 4 GHz CPU, 60 GB disk, DMA, 2 GB main memory and our own brand of system libraries consistently outperforms the state of the art.</td>
</tr>
</tbody>
</table>
Which abstract do you prefer?

Abstract (Take 1)
We provide a new algorithm that consistently outperforms the state of the art.

Abstract (Take 2)
We provide a new algorithm that on a Debian Linux machine with 4 GHz CPU, 60 GB disk, DMA, 2 GB main memory and our own brand of system libraries consistently outperforms the state of the art.

There are obvious, undisputed exceptions.
Making experiments parameterizable

This is huge
This is \textbf{huge}.

Parameters your code may depend on:
Making experiments parameterizable

This is huge
Parameters your code may depend on:

- credentials (OS, database, other)
Making experiments parameterizable

This is **huge**

Parameters your code may depend on:

- credentials (OS, database, other)
- values of important environment variables (usually one or two)
Making experiments parameterizable

This is huge

Parameters your code may depend on:

- credentials (OS, database, other)
- values of important environment variables (usually one or two)
- various paths and directories (see: environment variables)
This is huge
Parameters your code may depend on:
- credentials (OS, database, other)
- values of important environment variables (usually one or two)
- various paths and directories (see: environment variables)
- where the input comes from
Making experiments parameterizable

This is **huge**

Parameters your code may depend on:

- credentials (OS, database, other)
- values of important environment variables (usually one or two)
- various paths and directories (see: environment variables)
- where the input comes from
- switches (pre-process, optimize, prune, materialize, plot ...)

This is **huge**

Parameters your code may depend on:

- credentials (OS, database, other)
- values of important environment variables (usually one or two)
- various paths and directories (see: environment variables)
- where the input comes from
- switches (pre-process, optimize, prune, materialize, plot . . .)
- where the output goes
Purpose: have a very simple mean to obtain a test for the values

\[ f_1 = v_1, f_2 = v_2, \ldots, f_k = v_k \]
Making experiments parameterizable

Purpose: have a very simple mean to obtain a test for the values

\[ f_1 = v_1, f_2 = v_2, \ldots, f_k = v_k \]

Many tricks. Very simple ones:
Making experiments parameterizable

Purpose: have a very simple mean to obtain a test for the values

\[ f_1 = v_1, f_2 = v_2, \ldots, f_k = v_k \]

Many tricks. Very simple ones:

- `argc` / `argv`: specific to each class’ `main`
Purpose: have a very simple mean to obtain a test for the values

\[ f_1 = v_1, f_2 = v_2, \ldots, f_k = v_k \]

Many tricks. Very simple ones:

- `argc` / `argv`: specific to each class' `main`
- Configuration files
Making experiments parameterizable

Purpose: have a very simple mean to obtain a test for the values

\[ f_1 = v_1, f_2 = v_2, \ldots, f_k = v_k \]

Many tricks. Very simple ones:

- argc / argv: specific to each class' main
- Configuration files
- Java Properties pattern
Making experiments parameterizable

Purpose: have a very simple mean to obtain a test for the values

\[ f_1 = v_1, f_2 = v_2, \ldots, f_k = v_k \]

Many tricks. Very simple ones:

- \texttt{argc / argv}: specific to each class’ main
- Configuration files
- Java Properties pattern
- + command-line arguments
Making experiments parameterizable

Configuration files

Omnipresent in large-scale software

- Crucial if you hope for serious installations: see GNU software install procedure
- Decide on a specific relative directory, fix the syntax
- Report meaningful error if the configuration file is not found
Configuration files

Omnipresent in large-scale software

- Crucial if you hope for serious installations: see GNU software install procedure
- Decide on a specific relative directory, fix the syntax
- Report meaningful error if the configuration file is not found

**Pro:** human-readable even without running code
Making experiments parameterizable

Configuration files

Omnipresent in large-scale software

- Crucial if you hope for serious installations: see GNU software install procedure
- Decide on a specific relative directory, fix the syntax
- Report meaningful error if the configuration file is not found

Pro: human-readable even without running code
Con: the values are read when the process is created
Making experiments parameterizable

Java `util.Properties`

Flexible management of parameters for Java projects
Defaults + overriding

How does it go:

- Properties extends Hashtable
- Properties is a map of (key, value) string pairs
  
  ```
  {"dataDir", "./data"} {"doStore", "true"}
  ```

- Methods:
  
  - `getProperty(String s)`
  - `setProperty(String s1, String s2)`
  - `load(InputStream is)`
  - `store(OutputStream os, String comments)`
  - `loadFromXML(...), storeToXML(...)`
class Parameters{
    Properties prop;
    String[][] defaults = {{"dataDir", "/data"},
                           {"doStore", "true"}};

    void init(){
        prop = new Properties();
        for (int i = 0; i < defaults.length; i ++)
            prop.put(defaults[i][0], defaults[i][1]);
    }

    void set(String s, String v){ prop.put(s, v); }

    String get(String s){
        // error if prop is null!
        return prop.get(s);
    }
}
When the code starts, it calls Parameters.init(), loading the defaults.
The defaults may be overridden later from the code by calling set.
The properties are accessible to all the code.
The properties are stored in one place.
Simple serialization/deserialization mechanisms may be used instead of constant defaults.
Better init method

```java
class Parameters{
    Properties prop;
    ...
    void init(){
        prop = new Properties();
        for (int i = 0; i < defaults.length; i ++)
            prop.put(defaults[i][0], defaults[i][1]);
        Properties sysProps = System.getProperties();
        // copy sysProps into (over) prop!
    }
}
```

Call with: `java -DdataDir=./test -DdoStore=false pack.AnyClass`
Making your code parameterizable

The bottom line: you will want to run it in different settings

- With your or the competitor’s algorithm or special optimization
- On your desktop or your laptop
- With a local or remote MySQL server
- Make it easy to produce a point
- If it is very difficult to produce a new point, ask questions
Making your code parameterizable

The bottom line: you **will** want to run it in different settings
- With your or the competitor’s algorithm or special optimization
- On your desktop or your laptop
- With a local or remote MySQL server
- **Make it easy to produce a point**
- If it is very difficult to produce a new point, ask questions

You may omit coding like this:
The input data set files should be specified in source file:util.GlobalProperty.java.
Building a test suite

You already have:

- Designs
- Easy way to get any measure point

You need:

- Suited directory structure (e.g.: source, bin, data, res, graphs)
- Control loops to generate the points needed for each graph, under res/, and possibly to produce graphs under graphs
  - Even Java can be used for the control loops, but...
  - It does pay off to know how to write a loop in shell/perl etc.
Building a test suite

You already have:

- Designs
- Easy way to get any measure point

You need:

- Suited directory structure (e.g.: source, bin, data, res, graphs)
- Control loops to generate the points needed for each graph, under res/, and possibly to produce graphs under graphs
  - Even Java can be used for the control loops, but...
  - It does pay off to know how to write a loop in shell/perl etc.

You may omit coding like this:

Change the value of the 'delta' variable in distribution.DistFreeNode.java into 1,5,15,20 and so on.
Automatically generated graphs

You have:

- files containing numbers characterizing the parameter values and the results
- basic shell skills
Automatically generated graphs

You have:
- files containing numbers characterizing the parameter values and the results
- basic shell skills

You need: graphs

Most frequently used solutions:
- Based on Gnuplot
- Based on Excel or OpenOffice clone

Other solutions: R; Matlab (remember portability)
Automatically generating graphs with Gnuplot

Data file `results-m1-n5.csv`:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1234</td>
</tr>
<tr>
<td>2</td>
<td>2467</td>
</tr>
<tr>
<td>3</td>
<td>4623</td>
</tr>
</tbody>
</table>
Automatically generating graphs with Gnuplot

1. Data file `results-m1-n5.csv`:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1234</td>
</tr>
<tr>
<td>2</td>
<td>2467</td>
</tr>
<tr>
<td>3</td>
<td>4623</td>
</tr>
</tbody>
</table>

2. Gnuplot command file `plot-m1-n5.gnu` for plotting this graph:
Automatically generating graphs with Gnuplot

1. Data file `results-m1-n5.csv`:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1234</td>
</tr>
<tr>
<td>2</td>
<td>2467</td>
</tr>
<tr>
<td>3</td>
<td>4623</td>
</tr>
</tbody>
</table>

2. Gnuplot command file `plot-m1-n5.gnu` for plotting this graph:

```plaintext
set data style linespoints
set terminal postscript color
set output "results-m1-n5.eps"
set title "Execution time for various scale factors"
set xlabel "Scale factor"
set ylabel "Execution time (ms)"
plot "results-m1-n5.csv"
```
Automatically generating graphs with Gnuplot

1. Data file `results-m1-n5.csv`:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1234</td>
</tr>
<tr>
<td>2</td>
<td>2467</td>
</tr>
<tr>
<td>3</td>
<td>4623</td>
</tr>
</tbody>
</table>

2. Gnuplot command file `plot-m1-n5.gnu` for plotting this graph:

```
set data style linespoints
set terminal postscript color
set output "results-m1-n5.eps"
set title "Execution time for various scale factors"
set xlabel "Scale factor"
set ylabel "Execution time (ms)"
plot "results-m1-n5.csv"
```

3. Call `gnuplot plot-m1-n5.gnu`
Create an Excel file *results-m1-n5.xls* with the column labels:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Scale factor</td>
<td>Execution time</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
1. Create an Excel file `results-m1-n5.xlsx` with the column labels:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Scale factor</td>
<td>Execution time</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

2. Insert in the area B2-C3 a link to the file `results-m1-n5.csv`
Create an Excel file `results-m1-n5.xlsx` with the column labels:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Scale factor</td>
<td>Execution time</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Insert in the area B2-C3 a **link** to the file `results-m1-n5.csv`.

Create in the .xls file a graph out of the cells A1:B3, chose the layout, colors etc.
1. Create an Excel file `results-m1-n5.xls` with the column labels:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Scale factor</td>
<td>Execution time</td>
</tr>
<tr>
<td>2</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

2. Insert in the area B2-C3 a link to the file `results-m1-n5.csv`.

3. Create in the .xls file a graph out of the cells A1:B3, choose the layout, colors etc.

4. When the .csv file will be created, the graph is automatically filled in.
You may omit working like this:

In avgs.out, the first 15 lines correspond to $xyzT$, the next 15 lines correspond to $xYZT$, the next 15 lines correspond to $XyzT$, the next 15 lines correspond to $xyZT$, the next 15 lines correspond to $Xyzt$, the next 15 lines correspond to $XYZT$, and the next 15 lines correspond to $XyZT$. In each of these sets of 15, the numbers correspond to queries $1.1, 1.2, 1.3, 1.4, 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 3.3, 3.4, 4.1, 4.2, and 4.3$. 
You may omit working like this:

In avgs.out, the first 15 lines correspond to xyzT, the next 15 lines correspond to xYZT, the next 15 lines correspond to XyzT, the next 15 lines correspond to xyZT, the next 15 lines correspond to XyzT, the next 15 lines correspond to XyZT, and the next 15 lines correspond to XYZT. In each of these sets of 15, the numbers correspond to queries 1.1, 1.2, 1.3, 1.4, 2.1, 2.2, 2.3, 2.4, 3.1, 3.2, 3.3, 3.4, 4.1, 4.2, and 4.3.

... either because you want to do clean work, or because you don’t want this to happen:
Why you should take care to generate your own graphs

File `avgs.out` contains average times over three runs:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.666</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>12.3333</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>
Why you should take care to generate your own graphs

File `avgs.out` contains average times over three runs:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.666</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>12.3333</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

Copy-paste into OpenOffice 2.3.0-6.11-fc8:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13666</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>123333</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>
Why you should take care to generate your own graphs

File `avgs.out` contains average times over three runs:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.666</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>12.3333</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

Copy-paste into OpenOffice 2.3.0-6.11-fc8:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13666</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>123333</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

The graph doesn’t look good :-(

Manolescu, Manegold (INRIA, CWI)
Why you should take care to generate your own graphs

File `avgs.out` contains average times over three runs:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.666</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>12.3333</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

Copy-paste into OpenOffice 2.3.0-6.11-fc8:

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13666</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>123333</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

The graph doesn’t look good :-(
Hard to figure out when you have to produce by hand 20 such graphs and most of them look OK
Documenting your experiment suite

Very easy if they already portable, parameterizable, and if graphs are automatically generated
 Specify:

1. What the installation requires; how to install
2. For each experiment
   1. Extra installation if any
   2. Script to run
   3. Where to look for the graph
Documenting your experiment suite

Very easy if they already **portable, parameterizable, and if graphs are automatically generated**

Specify:

1. What the installation requires; how to install
2. For each experiment
   1. Extra installation if any
   2. Script to run
   3. Where to look for the graph
   4. How long it takes
Some numbers on how SIGMOD 2008 repeatability went

Accepted papers (78)

- All repeated
- Some repeated
- None repeated
- Excuse
- No submission
Some numbers on how SIGMOD 2008 repeatability went

Rejected verified papers (11)

- All repeated
- Some repeated
- None repeated
Some numbers on how SIGMOD 2008 repeatability went

- All verified papers (64)
  - All repeated
  - Some repeated
  - None repeated
Disclaimer

- We do *not* “blame” either the authors or the committee for anything
- We (tried to) anonymize and generalize the “war stories”
- Some of the war stories are unique, some occur more than once
Reasons not to provide the code for testing

Authors say

The work presented in this work heavily depends on the work of the primary author’s Ph.D. dissertation. The primary author has graduated and due to his job commitments is unable to spend enough time to get the code base together into an executable package. The project is coupled very tightly to other on-going research work and therefore require substantiate amount of time which the primary author does not have.
Authors say

(1) We use other people’s code and (2) we lost some old code. Due to the short notice, we could not write our own code/reproduce our lost code for these parts. If we have a 4 or 5 months ahead of the notice, we can give the code.
Authors say

(1) We use other people’s code and (2) we lost some old code. Due to the short notice, we could not write our own code/reproduce our lost code for these parts. If we have a 4 or 5 months ahead of the notice, we can give the code.

Authors say

This system has been in development for more than three years, and it is virtually impossible to package this system in a way that it can be run from the command line.
Reasons not to provide the code for testing

Authors say

We had to manually evaluate 300 queries that were chosen randomly, and determine if a result is relevant or not, based on our judgment. This was a tedious process that we assume your committee members do not want to repeat; in addition, different people have different judgment and achieving the same results is not feasible.
Reasons not to provide the code for testing

Authors say

We had to manually evaluate 300 queries that were chosen randomly, and determine if a result is relevant or not, based on our judgment. This was a tedious process that we assume your committee members do not want to repeat; in addition, different people have different judgment and achieving the same results is not feasible.

Authors say

The subsets were chosen randomly from a large dataset, and unfortunately no trace about the identity of the used documents has been kept. The experiments were performed months ago, and it wasn’t expected to send results to SIGMOD, that’s why we didn’t pay attention about keeping a trace.
Authors say

1) We can not create the batch files that reproduce the experiments in the requested format, and 2) the output of the simulator needs considerable work in order to be transformed according to the instructions, because it is based on prior work, and it is implemented before the SIGMOD instructions for the experimental evaluation. Our simulator does not take the input parameters from command line.
Encouragement from the authors

Authors say

This wasn’t too hard, and I think it was definitely worth it. We even found a mistake (thankfully a minor one, not affecting our conclusions) in our submission, so I think it was very helpful. Thanks a lot for taking the time to do the repeatability eval!
Encouragement from the authors

Authors say

This wasn’t too hard, and I think it was definitely worth it. We even found a mistake (thankfully a minor one, not affecting our conclusions) in our submission, so I think it was very helpful. Thanks a lot for taking the time to do the repeatability eval!

Authors say

It was helpful – we discovered an error in one of our graphs, for example, after the submission.
Encouragement from the authors

Authors say

I think the repeatability is very helpful, as we felt a great sense of achievement if other people can repeat our works and use our methods.
Encouragement from the authors

Authors say

I think the repeatability is very helpful, as we felt a great sense of achievement if other people can repeat our works and use our methods.

Authors say

I think in general it helps students to develop more solid software and algorithms although it involves work on both sides: our side to prepare more repeatable testing environment and solid test cases, and the review side to more testing and understand the method described in the paper.
Encouragements

Senior ACM SIGMOD officer says

I personally feel that this is a VERY important direction for SIGMOD to take leadership in. It is part of a natural maturing of the field. Up until now, we’ve been very lax in our experimentation, but this initiative gets everyone in the field thinking about it. Had this initiative been done for a minor workshop or conference, it would not have had much impact, but since it is done with one of the truly top conferences, I feel that everyone noticed, even those that eventually didn’t submit to that conference.

It is important to continue with this requirement. It will take literally years for the field to become comfortable with it and absorb it into its consciousness. At that point, every author doing an experiment will think instinctively about repeatability, which will raise our discipline to a new level of maturity.
It can be done

Repeatability reviewer says: one command...

- Built application from sources
- Ran all experiments successfully
- Produced all tables and graphs
- Re-built paper from sources, including the re-built tables and graphs
Longest(?) war story

- Experiments: Java programs that connect to standard RDBMS
- Instructions warned that data preparation for the full experiment might take more than 40 days(!) on a heavy 8x 3 GHz CPU server with 2 GB RAM & ≥300 GB (RAID?) disk (system)
- Authors & committee agreed to down-scale to 1/4 ⇒ 10 days for data preparation
- No info how RDBMS was (to be) configured/tuned
Longest(?) war story

- Evaluation machine: 4x dual-core Opteron @ 2 GHz, 16 GB RAM, 1 TB RAID-0 (4 disks)
- Default turned out to be (1) single threaded and (2) I/O bound, using only 200 MB of memory
- Committee tuned RDBMS to use all available memory and parallelized preparation task by distributing the workload over 4 clients
- Reduced preparation task from 10 days to 2 days
Two more preparation steps took another 4 days (couldn’t be parallelized)

4 experiments ran fine and finished within 4 hours in total

Last experiment failed as the provided version could not handle the reduced data set

Authors provided adapted version; however, simple re-run was no guaranteed to work as this experiment modified the database, and hence was not idempotent

Re-start from scratch required another 6 days to re-create the starting point for the last experiment.
Good and repeatable performance evaluation and experimental assessment require no fancy magic but rather solid craftsmanship.

Proper planning helps to keep you from “getting lost” and ensure repeatability.

Repeatable experiments simplify your own work (and help others to understand it better).

There is no single way how to do it right.

There are many ways how to do it wrong.

We provided some simple rules and guidelines what (not) to do.