What is a Good Benchmark?
Jeremy Arnold (IBM)
SPEC Power Architect
arnoldje@us.ibm.com
What is a Benchmark?

A program used to assess the performance of a computer system.
A program (or suite of programs) used to assess the performance of a computer system.
What is a Benchmark?

A program (or suite of programs) used to assess the performance of one or more computer systems.
What is a Benchmark?

A program (or suite of programs) used to assess the performance characteristics of one or more computer systems.
Why Do We Need Benchmarks?

To measure and study the behavior of a class of applications in a practical and repeatable way.
Benchmarks are used for a variety of purposes:

- Marketing
Benchmarks are used for a variety of purposes:

- Marketing
- Product Development
Benchmarks are used for a variety of purposes:

- Marketing
- Product Development
- Quality Assurance
Benchmarks are used for a variety of purposes:

- Marketing
- Product Development
- Quality Assurance
- Environment Verification
Benchmarks are used for a variety of purposes:

- Marketing
- Product Development
- Quality Assurance
- Environment Verification
- Research
Types of Benchmarks

Different uses call for different types of benchmarks:
Benchmarks can be assessed on several different characteristics, including:

- Relevance
- Reproducibility
- Fairness
- Verifiability
- Usability
Relevant benchmarks mimic the behavior of some class of real applications.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breadth</td>
<td>How large of a class of applications</td>
</tr>
<tr>
<td>Degree</td>
<td>How closely the behavior matches those applications</td>
</tr>
<tr>
<td>Scalability</td>
<td>Ability to use the resources of a wide range of systems</td>
</tr>
<tr>
<td>Environment</td>
<td>Measurements must be taken under realistic conditions</td>
</tr>
<tr>
<td>Variable Utilization</td>
<td>Energy efficiency varies at different utilizations</td>
</tr>
<tr>
<td>Multi-system</td>
<td>Energy sometimes can't be measured accurately for individual systems (e.g. blades)</td>
</tr>
</tbody>
</table>

Characteristics marked with ⚡ are mostly specific to energy-efficiency benchmarks.
Reproducibility

Benchmarks should produce results which can be reproduced by others.

<table>
<thead>
<tr>
<th>Consistency</th>
<th>Running the benchmark multiple times under the same conditions will produce the same results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>The hardware and software components and configuration are described in sufficient detail to allow an equivalent environment to be constructed</td>
</tr>
<tr>
<td>Power Measurements</td>
<td>Power should be measurable using a variety of devices</td>
</tr>
</tbody>
</table>
Systems can compete on their merits without artificial constraints.

<table>
<thead>
<tr>
<th>Portability</th>
<th>Benchmarks should run on any systems that is relevant for its target application space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credibility</td>
<td>Benchmarks are developed by a reputable organization (like SPEC), and not by a single vendor</td>
</tr>
<tr>
<td>Tuning</td>
<td>A balance between allowing reasonable tuning without “super-tuning” that wouldn't be appropriate for real applications</td>
</tr>
<tr>
<td>Fair Use</td>
<td>Benchmark rules may restrict the use of results to avoid misleading comparisons</td>
</tr>
<tr>
<td>Components</td>
<td>Which components of the system must have power measured?</td>
</tr>
</tbody>
</table>
Results can be verified to be accurate

<table>
<thead>
<tr>
<th><strong>Self-validating</strong></th>
<th>Automatic tests at runtime to confirm compliance with run rules</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tamper-resistant</strong></td>
<td>Detect manual modification of results</td>
</tr>
<tr>
<td><strong>Power Accuracy</strong></td>
<td>Accuracy of data from power analyzer depends on ranges and readings; requires dynamic verification</td>
</tr>
</tbody>
</table>
Easy-to-use benchmarks tend to have more results and better accuracy.

<table>
<thead>
<tr>
<th>Usability Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-describing</td>
<td>Includes tools for automatically discovery of system details</td>
</tr>
<tr>
<td>Practical</td>
<td>Runs on reasonably sized systems</td>
</tr>
<tr>
<td>Configurability</td>
<td>Allow flexibility for research</td>
</tr>
<tr>
<td>Energy Data Collection</td>
<td>Use of SPEC PTDaemon or other tools to automatically collect power data</td>
</tr>
</tbody>
</table>
Benchmark Characteristics

Benchmarks can be assessed on several different characteristics, including:

- Relevance
- Reproducibility
- Fairness
- Verifiability
- Usability
Components of a Benchmark

A benchmark is more than just an application:

- Workload
- Harness
- Reporting
- Documentation
- Run Rules
- Peer Review
- Fair Use Guidelines
What is a Good Benchmark?

Jeremy Arnold (IBM)
SPEC Power Architect
arnoldje@us.ibm.com