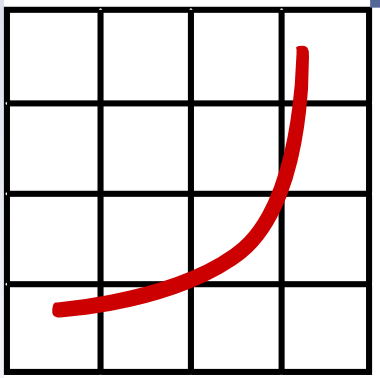
What is a Good Benchmark?

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



spec

What is a Benchmark?

A program
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 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

Benchmark Uses

Benchmarks are used for a variety of purposes:

- Marketing

Benchmark Uses

Benchmarks are used for a variety of purposes:

- Marketing
- Product Development

Benchmark Uses

Benchmarks are used for a variety of purposes:

- Marketing
- Product Development
- Quality Assurance

Benchmark Uses

Benchmarks are used for a variety of purposes:

- Marketing
- Product Development
- Quality Assurance
- Environment Verification

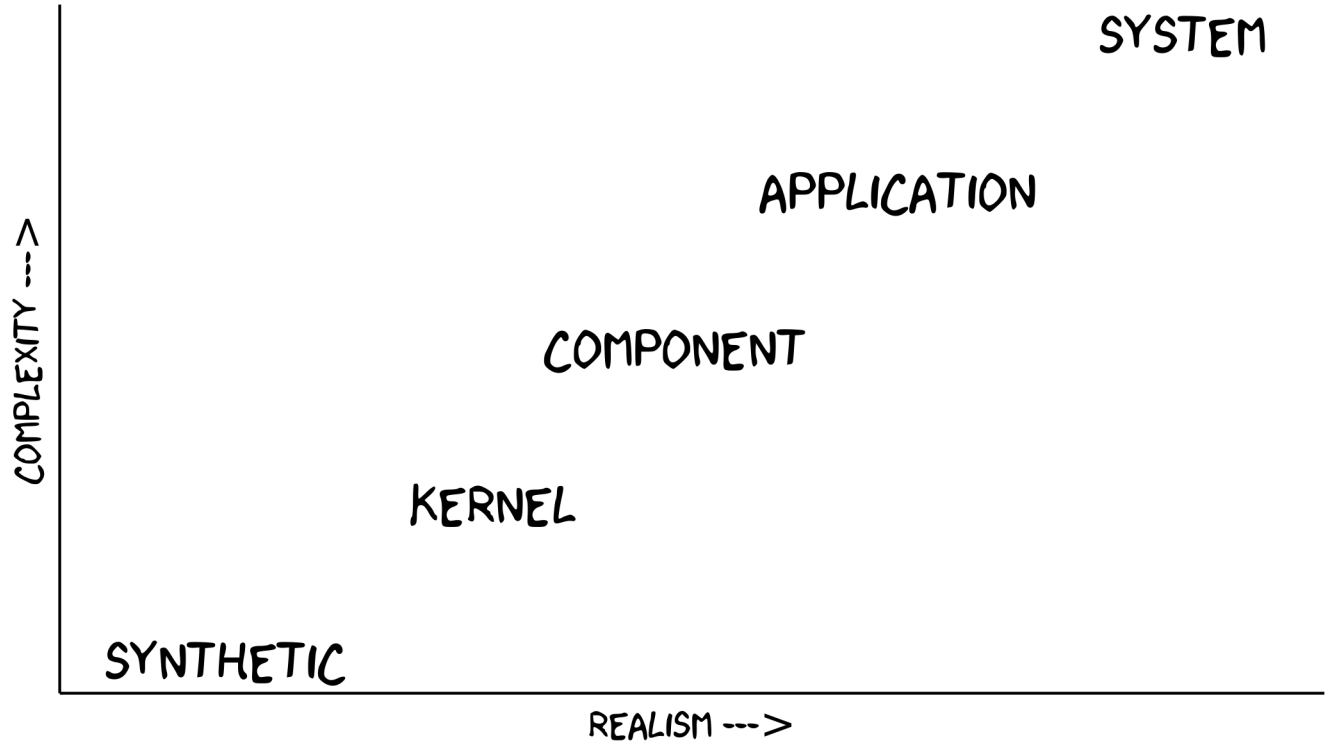
Benchmark Uses

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:



Benchmark Characteristics

Benchmarks can be assessed on several different characteristics, including:

- Relevance
- Reproducibility
- Fairness
- Verifiability
- Usability

Relevance

Relevant benchmarks mimic the behavior of some class of real applications.

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

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

Reproducibility

Benchmarks should produce results which can be reproduced by others.

Consistency	Running the benchmark multiple times under the same conditions will produce the same results
Description	The hardware and software components and configuration are described in sufficient detail to allow an equivalent environment to be constructed
Power Measurements 	Power should be measurable using a variety of devices

Fairness

Systems can compete on their merits without artificial constraints.

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

Verifiability

Results can be verified to be accurate

Self-validating	Automatic tests at runtime to confirm compliance with run rules
Tamper-resistent	Detect manual modification of results
Power Accuracy 	Accuracy of data from power analyzer depends on ranges and readings; requires dynamic verification

Usability

Easy-to-use benchmarks tend to have more results and better accuracy.

Self-describing	Includes tools for automatically discovery of system details
Practical	Runs on reasonably sized systems
Configurability	Allow flexibility for research
Energy Data Collection 	Use of SPEC PTDaemon or other tools to automatically collect power data

Benchmark Characteristics

Benchmarks can be assessed on several different characteristics, including:

- Relevance
- Reproducibility
- Fairness
- Verifiability
- Usability

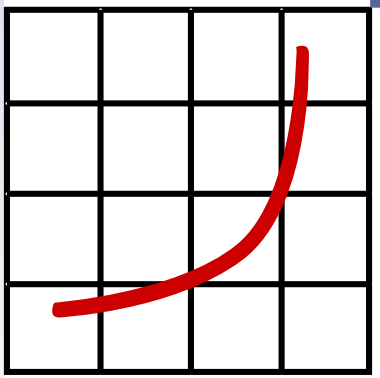
Components of a Benchmark

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



spec