Understanding Performance Metrics

Response Time vs. Throughput

When we say one computer is faster than another, what do we mean? Desktop users focus on **response time** (execution time) - the time between start and completion of a task. Warehouse-scale computer operators prioritize **throughput** - the total amount of work done in a given time.

Performance Relationship

$$n = \frac{ExecutionTime_Y}{ExecutionTime_X} = \frac{Performance_X}{Performance_Y}$$

When we say "X is n times faster than Y," we mean the execution time ratio equals n.

The only consistent and reliable measure of performance is the execution time of real programs. All alternatives have eventually led to misleading claims or design mistakes.

Types of Time Measurements

1

Wall-Clock Time

Also called response time or elapsed time, this is the total latency to complete a task, including disk accesses, memory accesses, I/O activities, and operating system overhead.

2

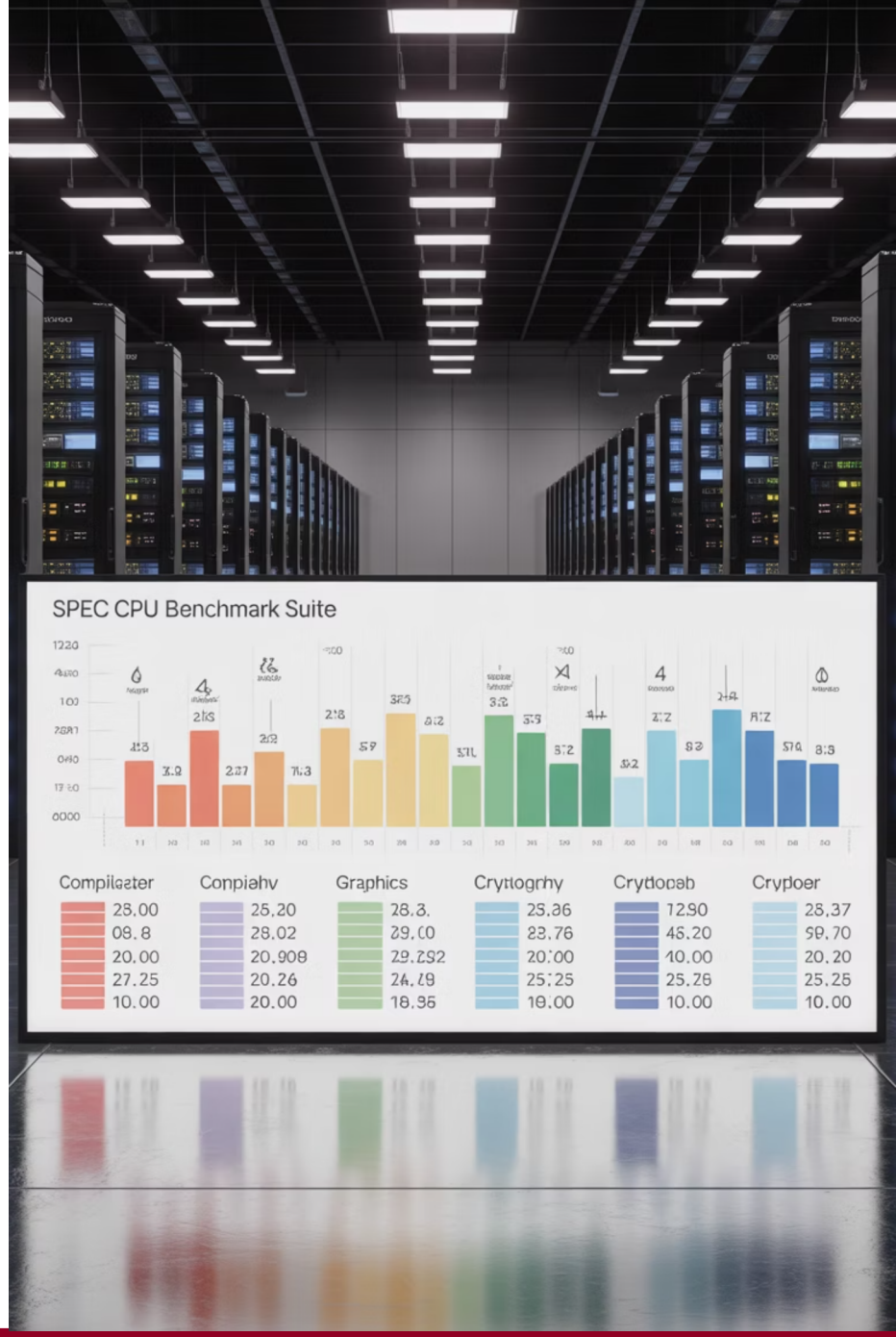
CPU Time

The time the processor spends computing for a specific program, not including time waiting for I/O or running other programs. This metric isolates processor performance.

For users running the same programs regularly, the best evaluation method would be comparing execution times of their workloads—the mixture of programs and commands they typically run. Most users must rely on benchmarks to predict performance.

Benchmark Suites

- To avoid placing too much emphasis on a single program, collections of benchmark applications called benchmark suites are used to measure processor performance across various applications. The weakness of any individual benchmark is lessened by the presence of others.
- The most successful standardized benchmark suite has been SPEC (Standard Performance Evaluation Corporation), which evolved from efforts in the late 1980s to deliver better workstation benchmarks. SPEC now covers many application classes, with results available at www.spec.org.



Server Benchmarks

1

Processor Throughput

SPECrate measures request-level parallelism by running multiple copies of SPEC CPU benchmarks simultaneously (usually one per processor).

2

I/O Performance

SPECSFS measures Network File System performance, testing both disk and network I/O as well as processor performance.

3

Web Server Performance

SPECWeb simulates multiple clients requesting both static and dynamic pages, plus clients posting data to the server.

4

Transaction Processing

TPC benchmarks measure the ability to handle transactions involving database accesses and updates, similar to airline reservations or bank ATM systems.

Summarizing Performance Results

The SPECRatio Approach

Rather than using weighted arithmetic means, SPEC normalizes execution times to a reference computer by dividing the reference time by the test computer time, yielding a ratio proportional to performance.

$$\frac{SPECRatio_A}{SPECRatio_B}$$

$$= \frac{ExecutionTime_B}{ExecutionTime_A} = \frac{Performance_A}{Performance_B}$$

Using Geometric Mean

Since SPECRatios are ratios rather than absolute execution times, the mean must be computed using the geometric mean:

$$GeometricMean = \sqrt[n]{\prod_{i=1}^{n} Sample_i}$$

This ensures that the ratio of geometric means equals the geometric mean of performance ratios, making the choice of reference computer irrelevant.

Example

Benchmarks	Ultra 5 time (sec)	Opteron time (sec)	SPECRatio	Itanium 2 time (sec)	SPECRatio	Opteron/Itanium times (sec)	Itanium/Opteron SPECRatios
wupwise	1600	51.5	31.06	56.1	28.53	0.92	0.92
swim	3100	125.0	24.73	70.7	43.85	1.77	1.77
mgrid	1800	98.0	18.37	65.8	27.36	1.49	1.49
applu	2100	94.0	22.34	50.9	41.25	1.85	1.85
mesa	1400	64.6	21.69	108.0	12.99	0.60	0.60
galgel	2900	86.4	33.57	40.0	72.47	2.16	2.16
art	2600	92.4	28.13	21.0	123.67	4.40	4.40
equake	1300	72.6	17.92	36.3	35.78	2.00	2.00
facerec	1900	73.6	25.80	86.9	21.86	0.85	0.85
ammp	2200	136.0	16.14	132.0	16.63	1.03	1.03
lucas	2000	88.8	22.52	107.0	18.76	0.83	0.83
fma3d	2100	120.0	17.48	131.0	16.09	0.92	0.92
sixtrack	1100	123.0	8.95	68.8	15.99	1.79	1.79
apsi	2600	150.0	17.36	231.0	11.27	0.65	0.65
Geometric mean			20.86		27.12	1.30	1.30

- SPECfp2000 execution times (in seconds) for the Sun Ultra 5—the reference computer of SPEC2000—and execution times and SPECRatios for the AMD Opteron and Intel Itanium 2. (SPEC2000 multiplies the ratio of execution times by 100 to remove the decimal point from the result, so 20.86 is reported as 2086.)