Forgotten Aspects of Computer Performance

John Gustafson
Ames Lab-ISU
Ames, Iowa

This work was supported by the Applied Mathematical Sciences Program of the USDOE, contract W-7405-ENG-82
Our community sometimes equates “performance” with “teraflops” or “speedup,” and forgets or takes for granted other aspects of performance.

It is time to reconsider those other aspects. If we don’t, we could wind up with 100 teraflops and yet have low performance!
Typical Marketing Chart

What They Don’t Tell You
High-performance computing cannot ignore conversion effort when comparing two approaches.

Ease-of-Use is Part of Performance
What is *Speed*?

Sure, speed = work/time
But what is *work*, for a computer?
Higher speed isn’t just time reduction.

*From Ambrose Bierce, The Devil’s Dictionary*:

**Logic, n.** The art of thinking and reasoning in strict accordance with the limitations and incapacities of human misunderstanding. The basis of logic is the syllogism, consisting of a major and a minor premise and a conclusion—thus:

- **Major Premise**: Sixty men can do a piece of work sixty times as quickly as one man.
- **Minor Premise**: One man can dig a post-hole in sixty seconds.
- **Conclusion**: Sixty men can dig a post-hole in one second.

This may be called the syllogism arithmetical, in which, by combining logic and mathematics, we obtain a double certainty, and are twice blessed.
Assessing Computational “Work”

PARAMETER (N = 1000, N2 = N \times N, N3 = N2 \times N)
REAL A(N), B(N)
INTEGER L(N3), M(N3)

DO 1 I = 1, N
   A(I) = B(I) \times B(I)
1

DO 2 I = 1, N2
   L(I) = N2 - I
2

DO 3 I = 1, N3
   M(I) = L(I)
3

END
Can operations be standardized and counted?

What does the “64-bit” arithmetic mean?
- Mantissas range from 47 to 56 bits
- Even IEEE format doesn’t prevent differences.
- How much work is, say $x \times y$? Or $\text{ABS}(x)$?

Count fetches and stores as operations?
- Is there error detection/correction?
- Out of what size address space?
- Word aligned? In cache? Pipelined? Interleaved? ...

The work is not the “operation” count
0% finance charge till March

- 100 MHz!
- 8 MB RAM
- 1.2 GB hard drive
- Quad-speed CD-ROM drive

1699.88
Packard Bell multimedia computer with 100 MHz Pentium® processor

- Monitor included
- 8 MB RAM
- 10 GB hard drive
- Quad-speed CD-ROM drive

1999.99
Apple® Macintosh® Performa® 6200 multimedia computer with 75 MHz Power PC 603 microprocessor

See inside back cover for important 9% finance charge.
The LINPACK Myth:

“If a computer has a higher rank on the LINPACK benchmark, it will run my application faster.”

- LINPACK needs very little memory
- LINPACK needs little memory bandwidth
- LINPACK tolerates high latency
- Kernel: 67% Floating-Point Math, 33% Load/Store

Real application instruction frequencies:

<table>
<thead>
<tr>
<th>Instruction Type</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Load/Store</td>
<td>84%</td>
</tr>
<tr>
<td>Bit Manipulations</td>
<td>7%</td>
</tr>
<tr>
<td>Floating-Point Math</td>
<td>4%</td>
</tr>
<tr>
<td>Call/Return</td>
<td>3%</td>
</tr>
<tr>
<td>System Management</td>
<td>2%</td>
</tr>
<tr>
<td>Character Operations</td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>

Source: “IBM RT PC Computer Technology,” IBM Form No. SA23-1057, 1986
Things to Do With Unbalanced Systems

- No MB/sec
  - Matrix Multiply
  - LAPACK
  - Radar Cross-Section

- No MFLOPS
  - Sorting
  - Transaction Processing
  - Word Processing

- No Communication
  - Mandelbrot, Fractals
  - Monte Carlo
  - Ray Tracing
**Peak FLOPS No Longer Predicts Actual FLOPS**

<table>
<thead>
<tr>
<th>Application MFLOPS</th>
<th>TMC CM-2</th>
<th>Cray Y-MP/8</th>
<th>Intel iPSC/860</th>
<th>nCUBE nCUBE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Peak&quot; rated MFLOPS</td>
<td>1104</td>
<td>436</td>
<td>362</td>
<td>2605</td>
</tr>
<tr>
<td>Fraction of Peak</td>
<td>41%</td>
<td>3%</td>
<td>5%</td>
<td>106%</td>
</tr>
</tbody>
</table>

Bar chart: Application MFLOPS vs. Fraction of Peak
Means-Based vs Ends-Based Metrics

**MEANS-BASED**
- Flop/s
- Bytes of RAM
- Number of Processors
- Use of Commodity Parts
- Word Size
- ECC Memory
- Speedup

**ENDS-BASED**
- Time to Compute Answer
- Detail, Content of Answer
- Feasible Problems to Attempt
- Cost, Availability of System
- Closeness to Actual Physics
- Reliability of Answer
- Product Line Range
Which Algorithm Would You Pick?

Explicit Timestepping ↔ Implicit Timestepping
Conventional Matrix Multiply ↔ Strassen, Winograd Methods
Cholesky Decomposition ↔ PC Conjugate Gradient
All-to-All N-Body Methods ↔ Barnes-Hut, Greengard
Successive Over-Relaxation ↔ Multigrid
Time-Domain Operators ↔ FFT's
Recompute Gaussian Integrals ↔ Compute Once and Store
Material Property Function ↔ Table Look-Up

HIGHER FLOP/S RATES

FASTER ANSWERS
Example: Dense vs Sparse Matrix Multiply

```
DO 1 K=1,1000
  DO 1 J=1,1000
    DO 1 I=1,1000
      C(I,J) = C(I,J) + A(I,K) * B(K,J)
  END IF
1 CONTINUE

DO 2 K=1,1000
  DO 2 J=1,1000
    IF (B(K,J).NE.0.) THEN
      DO 1 I=1,1000
        IF (A(I,K).NE.0.) C(I,J) = ...
      END IF
1 END IF
2 CONTINUE
```

Cray Food: 
>80% Peak

Ruins Pipelining

2.0 GFLOPS 1.0 Seconds

0.2 GFLOPS 0.1 Seconds

Ten Times Faster or Ten Times Slower?
N-Body Challenges

- Planetary position was a Grand Challenge in the 1940s.
- Size of $N$ often taken as figure of merit. “Billion-particle simulation.”
- GRAPE processor project uses all-to-all method, measures ops/sec.
- Materials science, astrophysics, and fusion all require N-body variants.
- Greengard, Barnes-Hut et al. made force calculation take $O(Np^2)$ work

To double the physical accuracy of any N-body method appears to take at least four times as much work. This is not generally understood. Physical accuracy and $N$ are not proportional.
Hmm... looks suspicious

Ah. Much better. Looks right.

The “Eyeball Metric”
Only Works for Graphics
History of Concern for Validity

1940  Ignored. Just use lots of decimals.
1950  Monte Carlo debated; roundoff studied.
1970  First 60-bit, 64-bit computer architectures.
1980  PASCAL-SC, ACRITH, ULTRITH.

Comparison with physical experiments is getting rarer. Accuracy is neglected; speedup, FLOPS emphasized. Different answers for parallel methods cause surprise.
Example: LINPACK Residuals

Value for n (maximum = 1160):
Please send the results of this run to:

Jack J. Dongarra
Computer Science Department
University of Tennessee
Knoxville, Tennessee 37996-1300

Fax: 615-974-8296
Internet: dongarra@cs.utk.edu

\[
\begin{array}{cccccc}
\text{norm. resid} & \text{resid} & \text{mache}p & x(1) & x(n) \\
1.33497627E+01 & 2.96423996E-12 & 2.22044605E-16 & 1.00000000E+00 & 1.00000000E+00 \\
\end{array}
\]

Which of these gets the most attention?

times are reported for matrices of order 1000
factor solve total mflops unit ratio
times for array with leading dimension of 580
2.415E+01 1.844E+00 2.600E+01 2.572E+01 7.776E-02 4.642E+02
Measuring Answer Quality

If $F$ is the answer, bound it rigorously by $F^+$ and $F^-$. Define total error as

$$E = \int \int \int \int (F^+ - F^-) \, dx \, dy \, dz \, dt$$

and define the answer quality $Q = 1/E$.

This has several desirable consequences:

- Removes need for flops/second or instructions/second metrics
- Allows fair comparison of different algorithms and computer architectures
- Permits clear and rigorous statement of goal for Grand Challenges
One Approach: Integral Equations

One way to find $F^+$ and $F^-$ is to restate the PDE as an integral equation, if possible. Integral equations of the Second Kind are usually tractable.

$$f(x) - \int K(x, s) f(s) \, ds = g(x)$$

One can bound each variable on a discretization, and bound the integral. Physical reasoning may be needed to get an initial bound.

We have found quality definitions and corresponding algorithms for

- Nonlinear ODEs
- Heat transfer problems
- The N-body problem
- Laplace’s equation
Computational Science as a Complement to Traditional Science

Experimental (chemists, physicists, engineers)

Interactions Between Experiment Theory, and Computation

Theoretical (chemists, physicists, mathematicians)

Computational (computer scientists, mathematicians, engineers)

↑ Model real process
↑ Suggest experiments
↑ Analyze data
↑ Control apparatus
↓ Generate data
Suggest theory ↓
Test theory ↓
Interpret experiments ↑
Suggest experiments ↑
Provide equations →
Interpret results →
← Suggest theory
← Perform accurate calculations
← Perform large-scale calculations

Experimental (chemists, physicists, engineers)

Suggest theory
Test theory
Generate data

Interactions Between Experiment Theory, and Computation

Theoretical (chemists, physicists, mathematicians)

Computational (computer scientists, mathematicians, engineers)

→ Interpret experiments
Suggest experiments
Provide equations

Experimental (chemists, physicists, engineers)

Suggest theory
Test theory
Generate data

Interactions Between Experiment Theory, and Computation

Theoretical (chemists, physicists, mathematicians)

Computational (computer scientists, mathematicians, engineers)

→ Interpret experiments
Suggest experiments
Provide equations

Experimental (chemists, physicists, engineers)

Suggest theory
Test theory
Generate data

Interactions Between Experiment Theory, and Computation

Theoretical (chemists, physicists, mathematicians)

Computational (computer scientists, mathematicians, engineers)

→ Interpret experiments
Suggest experiments
Provide equations

Experimental (chemists, physicists, engineers)

Suggest theory
Test theory
Generate data

Interactions Between Experiment Theory, and Computation

Theoretical (chemists, physicists, mathematicians)

Computational (computer scientists, mathematicians, engineers)

→ Interpret experiments
Suggest experiments
Provide equations
What Happens If There is No Experimental Component?

Theoretical (chemists, physicists, mathematicians)

- Provide equations →
- Interpret results →

Computational (computer scientists, mathematicians, engineers)

← Suggest theory
← Perform accurate calculations
← Perform large-scale calculations

“More teraflops” isn’t the answer.
ASCI Scenario 1: (cost of about $100M)
- Parallel computer rated at 3 TFLOPS sustains 1 TFLOPS, 70% parallel efficiency modeling nuclear weapon test.
- No proof of correctness, no accuracy goal.

ASCI Scenario 2: (cost of about $0.01M)
- Computer rated at 0.0002 TFLOPS sustains unknown TFLOPS and efficiency modeling nuclear weapon test.
- Answers have 95% confidence, match prior physical experiments.

If the Answer if Wrong, the Performance is ZERO.
Reliability Scenario 1:

Parallel supercomputer rated at 1 TFLOPS has Mean Time Between Failure of 4 hours.
Mean Job Size < $10^{16}$ f.p. operations.

Reliability Scenario 2:

Parallel fault-tolerant system rated at 0.01 TFLOPS has MTBF of 10 months.
Mean Job Size < $10^{17}$ f.p. operations.

If the Answer is Wrong, the Performance is ZERO.
Even with infinitely many processors, speedup can never exceed $1/s$.

\[
\text{Serial Processor Time: } T = s \\
\text{Hypothetical Parallel Processor Time: } T = s + \frac{p}{N}
\]
Why I Hate Amdahl’s Law

- Problems grow as computer power increases; the time people are willing to wait for an answer is relatively constant.

- Uniprocessor problems use only 1/N of the memory on N distributed memory processors.

- Memory gets closer with more processors.

- Computers don’t double in cost when the number of processors doubles.

- The parallel array can be shared.
Human Factors Limit Problem Size.
Financial Limits on Performance

DILBERT

THE FINANCE DEPARTMENT HAS ANALYZED YOUR COMPUTING NEEDS AND DECIDED TO GIVE YOU A 286 PC

THAT SHOULD BE SUFFICIENT FOR THE 3D-RENDERING YOU NEED TO DO

BESIDES, HOW MANY TIMES ARE YOU GOING TO DO 3D-RENDERING IN YOUR CAREER?

ONCE, IF I HURRY
“Ptolemaic” Speedup Theory

The retrograde motion of Saturn can be explained by building circles on circles, preserving the theory that the Earth is fixed at the center of the Universe.

Scaled speedup can be explained by making $s$ a function of $N$, preserving Amdahl’s theory.

\[
\text{Speedup} = \frac{1}{\left(\frac{s}{s+pN}\right) + \left(1-\frac{s}{s+pN}\right) \frac{N}{N}} = N + (1-N)s'
\]
Memory is Never Flat.
Speed gets used to Increase Problem Size, Not to Reduce Execution Time.
When the problem can scale to match available power, there is no limit to speedup.

\[
\text{Hypothetical Serial Processor Time} \quad \text{Time} = s' + Np' \\
\text{Parallel Processor Time} \quad \text{Time} = 1
\]

\[
\text{Scaled Model:} \quad \text{Speedup} = s' + Np' = N + (1 - N)s'
\]
Visualizing “Speed”

Miles per Hour

Time

0.1 sec 1 sec 10 sec 1.7 min 17 min 2.8 hrs 1.2 days

0.1

1

10

100

1000

Human on foot

Car

Commercial Jet
A New Way to Graph “Speedup”

- 64 Processors
- 32 Processors
- 16 Processors
- 8 Processors
- 4 Processors
- 2 Processors
- Serial Version

Speed in simulated Photons/sec

Time in Seconds

Speedup
When Power and Problem Size Don’t Match

Problem: What is 7 times 13?

<table>
<thead>
<tr>
<th>Computer</th>
<th>Time to Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>2 seconds</td>
</tr>
<tr>
<td>Pocket Calculator</td>
<td>30 seconds</td>
</tr>
<tr>
<td>RISC Workstation</td>
<td>5 minutes</td>
</tr>
<tr>
<td>Cray C90</td>
<td>4 hours</td>
</tr>
</tbody>
</table>
If you use too big a computer, you’ll take more time. Not less.

Big Computers are for Big Problems.
The Human Element in Performance Evaluation

ACTUAL TIME

TYPICAL BENCHMARK
“I was a computer performance measurement specialist. What are you here for?”