Recap and Plan

- Scaled speedup: operate near the memory boundary
- Memory systems on modern processors are complicated.
- The performance of a simple program can depend on the details of the micro-architecture.
- Today we will study matrix multiplication optimizations
  - An important kernel in some scientific problems
  - Closely related to other algorithms, e.g., transitive closure on a graph using Floyd-Warshall
  - Optimization ideas can be used in other problems
  - The best case for optimization payoffs
  - The most well-studied algorithm in high performance computing

Outline

- Performance Modeling
- Matrix-Vector Multiply (Warmup)
- Matrix Multiply Cache Optimizations
- Bag of Tricks
- Research in Matrix Multiply

Using a Simple Model of Memory to Optimize

- Assume just 2 levels in the hierarchy, fast and slow
- All data initially in slow memory
  - $m$ = number of memory elements (words) moved between fast and slow memory
  - $t_m$ = time per slow memory operation
  - $f$ = number of arithmetic operations
  - $t_f$ = time per arithmetic operation
- Minimum possible time = $f \cdot t_f$ when all data in fast memory
- Actual time
  - $f \cdot t_f + m \cdot t_m = f \cdot t_f \cdot \left(1 + \frac{t_m}{t_f} \cdot \frac{1}{q}\right)$
  - Larger $q$ means time closer to minimum $f \cdot t_f$

Warm up: Matrix-vector multiplication

\[
\text{[implements } y = y + A \cdot x]\]
\[
\text{for } i = 1:n \text{ for } j = 1:n
\]
\[
y(i) = y(i) + A(i,j) \cdot x(j)
\]

- $m$ = number of slow memory refs = $3n + n^2$
- $f$ = number of arithmetic operations = $2n^2$
- $q = \frac{f}{m} \approx 2$
- Matrix-vector multiplication limited by slow memory speed
Modeling Matrix-Vector Multiplication

- Compute time for \(nxn = 1000 \times 1000\) matrix
- Time
  \[T = t_f + t_m = f \cdot t_f + m \cdot t_m = f \cdot t_f (1 + \frac{m}{t_f} \cdot \frac{1}{q})\]
  \[= 2 \cdot n^2 \cdot (1 + 0.5 \cdot \frac{t_m}{t_f})\]
- For \(t_f\) and \(t_m\), using data from R. Vuduc’s PhD (pp 352-3)
- For \(t_m\), use words-per-cache-line / minimum-memory-latency

<table>
<thead>
<tr>
<th>Machine</th>
<th>Clock (MHz)</th>
<th>Peak Mflop/s</th>
<th>Mem Lat (Min, Max)</th>
<th>Linesize (Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ultra 2i</td>
<td>333</td>
<td>667</td>
<td>36, 60</td>
<td>24.8</td>
</tr>
<tr>
<td>Ultra 3</td>
<td>900</td>
<td>1800</td>
<td>26, 200</td>
<td>32</td>
</tr>
<tr>
<td>Pentium 3</td>
<td>500</td>
<td>1000</td>
<td>26, 60</td>
<td>32</td>
</tr>
<tr>
<td>Pentium 3MHz</td>
<td>800</td>
<td>800</td>
<td>40, 60</td>
<td>32</td>
</tr>
<tr>
<td>PowerPC 603</td>
<td>370</td>
<td>1500</td>
<td>35, 1000</td>
<td>128</td>
</tr>
<tr>
<td>PowerPC 750</td>
<td>1300</td>
<td>5200</td>
<td>60, 10000</td>
<td>128</td>
</tr>
<tr>
<td>Power4</td>
<td>800</td>
<td>3200</td>
<td>36, 85</td>
<td>32</td>
</tr>
<tr>
<td>Itanium1</td>
<td>800</td>
<td>9600</td>
<td>11, 60</td>
<td>24.8</td>
</tr>
<tr>
<td>Itanium2</td>
<td>1000</td>
<td>12200</td>
<td>11, 60</td>
<td>24.8</td>
</tr>
</tbody>
</table>

What Simplifying Assumptions

- What simplifying assumptions did we make in this analysis?
  - Ignored parallelism in processor between memory and arithmetic within the processor
  - Sometimes drop arithmetic term in this type of analysis
  - Assumed fast memory was large enough to hold three vectors
    - Reasonable if we are talking about any level of cache
    - Not necessarily if we are talking about registers (~32 words)
  - Assumed cost of a fast memory access is 0
    - Reasonable if we are talking about registers
    - Not necessarily if we are talking about cache (1-2 cycles for L1)
  - Memory latency is constant
    - Could simplify even further by ignoring memory operations in X and Y vectors
  - Mflop rate/element = \(\frac{2}{2 \cdot t_f + t_m}\)

Validating the Model

- How well does the model predict actual performance?
  - Using DGEMV: Most highly optimized code for the platform
  - Model sufficient to compare across machines
  - But under-predicting on most recent ones due to latency estimate

\[\text{Actual} < \text{Model} \times \text{Over-Prediction Rate}\]

“Naïve” Matrix Multiply

\[C(i,j) = C(i,j) + A(i,k) \times B(k,j)\]

Algorithm has \(2 \times n^3 = O(n^3)\) Flops and operates on \(3 \times n^2\) words of memory

Matrix Multiply on RS/6000

\[T = N^2 \text{ cycles/flop}\]

Size 2000 took 5 days

\(O(N^3)\) performance would have constant cycles/flop

Performance looks much closer to \(O(N^2)\)

Note on Matrix Storage

- A matrix is a 2-D array of elements, but memory addresses are “1-D”
- Conventions for matrix layout
  - by column, or “column major” (Fortran default): \(A[i,j]\) at \(A + i + j \times n\)
  - by row, or “row major” (C default): \(A[i,j]\) at \(A + i \times n + j\)

<table>
<thead>
<tr>
<th>Column major</th>
<th>Row major</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column major in memory</td>
<td></td>
</tr>
<tr>
<td>cachelines</td>
<td></td>
</tr>
<tr>
<td>Blue row of matrix is stored in red cachelines</td>
<td></td>
</tr>
</tbody>
</table>

Slide source: Larry Carter, UCSD
"Naive" Matrix Multiply

```plaintext
[implements C = C + A*B]
for i = 1 to n
  for j = 1 to n
    for k = 1 to n
      C(i,j) = C(i,j) + A(i,k) * B(k,j)
```

Reuse value from a register

Stride-N access to one row

Sequential access through entire matrix

- When cache (or TLB or memory) can't hold entire B matrix, there will be a miss on every line.
- When cache (or TLB or memory) can't hold a row of A, there will be a miss on each access

*Assumes column-major order

---

Matrix Multiply on RS/6000

- Cache miss every 16 iterations
- Page miss every 512 iterations

---

"Naive" Matrix Multiply

Number of slow memory references on unblocked matrix multiply

```plaintext
m = n\(^3\) read each column of B \(n\) times
  + \(n\) read each row of A \(n\) times
  + 2\(n\) read and write each element of C \(n\) times

So q = f / m = 2\(n\) \(n + 3\) for large \(n\), no improvement over matrix-vector multiply
```

---

Blocked (Tiled) Matrix Multiply

Consider A, B, C to be \(N\) by \(N\) matrices of \(b\) by \(b\) subblocks where \(b=n / N\) is called the block size

```plaintext
for i = 1 to \(N\)
  for j = 1 to \(N\)
    (read block C(i,j) into fast memory)
    for k = 1 to \(N\)
      (read block A(i,k) into fast memory)
      (read block B(k,j) into fast memory)
      C(i,j) = C(i,j) + A(i,k) * B(k,j) (do a matrix multiply on blocks)
      (write block C(i,j) back to slow memory)
```

Recall:

- \(m\) is amount memory traffic between slow and fast memory
- \(f\) is number of floating point operations, \(2n^3\) for this problem

So computational intensity q = f / m = \(2n\) / \((2N + 2) * n^2\) for large \(n\)

- Can be much faster than matrix-vector multiply (q=2)
Using Analysis to Understand Machines

The blocked algorithm has computational intensity \( q \approx b \)

- The larger the block size, the more efficient our algorithm will be
- Limit: All three blocks from A,B,C must fit in fast memory (cache), so we cannot make these blocks arbitrarily large
- Assume your fast memory has size \( M_{\text{fast}} \)
  \[ 3b^2 \leq M_{\text{fast}} \quad \text{so} \quad q \approx b \leq \sqrt{M_{\text{fast}}/3} \]

- To build a machine to run matrix multiply at the peak arithmetic speed of the machine, we need a fast memory of size
  \[ M_{\text{fast}} \geq 3b^2 = 3q^2 = 3(T_m/T_f)^2 \]

- There is a lower bound result that says we cannot do any better than this (using only algebraic associativity)

  Theorem (Hong & Kung, 1981): Any reorganization of this algorithm (that uses only algebraic associativity) is limited to \( q = O(\sqrt{M_{\text{fast}}}) \)

Limits to Optimizing Matrix Multiply

- The blocked algorithm changes the order in which values are accumulated into each \( C[i,j] \) by applying associativity
- The previous analysis showed that the blocked algorithm has computational intensity:
  \[ q \approx b \leq \sqrt{M_{\text{fast}}/3} \]

  • There is a lower bound result that says we cannot do any better than this (using only algebraic associativity)
  • Theorem (Hong & Kung, 1981): Any reorganization of this algorithm (that uses only algebraic associativity) is limited to \( q = O(\sqrt{M_{\text{fast}}}) \)

Basic Linear Algebra Subroutines

- Industry standard interface (evolving)
- Vendors, others supply optimized implementations

  • History
    - BLAS1 (1970s):
      - vector operations: dot product, saxpy (\( y=\alpha x+y \)), etc
    - \( m=2n, f=2n, q \approx 1 \) or less
    - BLAS2 (mid 1980s)
      - matrix-vector operations: matrix vector multiply, etc
      - \( m=n^2, f=2n^2, q \approx 2, \) less overhead
      - somewhat faster than BLAS1
    - BLAS3 (late 1980s)
      - matrix-matrix operations: matrix matrix multiply, etc
      - \( m \approx 4n^2, f=O(n^3), \) so \( q \) can possibly be as large as \( n \), so BLAS3 is potentially much faster than BLAS2
  - Good algorithms used BLAS3 when possible (LAPACK)

BLAS speeds on an IBM RS6000/590

Search Over Block Sizes

- Performance models are useful for high level algorithms
- Helps in developing a blocked algorithm
- Models have not proven very useful for block size selection
  - too complicated to be useful
  - See work by Sid Chatterjee for detailed model
- too simple to be accurate
  - Multiple multidimensional arrays, virtual memory, etc.
- Some systems use search
  - Atlas – being incorporated into Matlab
  - BeBOP – http://www.cs.berkeley.edu/~richie/bebop

What the Search Space Looks Like

A 2-D slice of a 3-D register-tile search space. The dark blue region was pruned.
(Platform: Sun Ultra-III, 333 MHz, 667 Mflop/s peak, Sun cc v5.0 compiler)
ATLAS (DGEMM n = 500)

- ATLAS is faster than all other portable BLAS implementations and it is comparable with machine-specific libraries provided by the vendor.

**Source:** Jack Dongarra

Tiling Alone Might Not Be Enough

- Naive and a "naïvely tiled" code on Itanium 2
- Searched all block sizes to find best, b=56
- Starting point for hw1

**Architectures**

AMD Athlon-600
DEC ev56-533
DEC ev6-500
HP9000/735/135
IBM PPC604-112
IBM Power2-160
IBM Power3-200
Pentium Pro-200
Pentium II-266
Pentium III-550
SGI R10000ip28-200
SGI R12000ip30-270
Sun UltraSparc2-200

**MFLOPS**

| Vendor BLAS | ATLAS BLAS | F77 BLAS |

**Source:** Jack Dongarra

Tiling for registers
- Loop unrolling, use of named "register" variables
- Tiling for multiple levels of cache and TLB
- Exploiting fine-grained parallelism in processor
  - Superscalar, pipelining
- Complicated compiler interactions
- Hard to do by hand (but you’ll try)
- Automatic optimization an active research area
  - SheBOP: shebop.cs.berkeley.edu/
  - PHiPAC: www.josu.berkeley.edu/~bilmes/phipac
  - In particular tr-98-035.ps.gz
- ATLAS: www.netlib.org/atlas

**Optimizing in Practice**

- Tiling for multiple levels of cache and TLB
- Exploiting fine-grained parallelism in processor
- Superscalar, pipelining
- Complicated compiler interactions
- Hard to do by hand (but you’ll try)
- Automatic optimization an active research area
  - SheBOP: shebop.cs.berkeley.edu/
  - PHiPAC: www.josu.berkeley.edu/~bilmes/phipac
  - In particular tr-98-035.ps.gz
  - ATLAS: www.netlib.org/atlas

**Removing False Dependencies**

- Using local variables, reorder operations to remove false dependencies

```
float f1 = b[i];
float f2 = b[i+1];
```

With some compilers, you can declare a and b unaliased.
- Done via "restrict pointers," compiler flag, or pragma

Exploit Multiple Registers

- Reduce demands on memory bandwidth by pre-loading into local variables

```
while( ... ) {
  *res++ = filter[0]*signal[0]
  + filter[1]*signal[1]
  + filter[2]*signal[2];
  signal++;
```

**Example is a convolution**

Minimize Pointer Updates

- Replace pointer updates for strided memory addressing with constant array offsets

```
f0 = r0[0]; r0 += 4;
f1 = r0[4]; r0 += 4;
f2 = r0[8]; r0 += 4;
```

**Pointer vs. array expression costs may differ.**
- Some compilers do a better job at analyzing one than the other
Loop Unrolling

- Expose instruction-level parallelism

```java
float f0 = filter[0], f1 = filter[1], f2 = filter[2];
float s0 = signal[0], s1 = signal[1], s2 = signal[2];
float res = 0;
do {
    signal += 3;
    s0 = signal[0];
    res[0] = f0*s0 + f1*s1 + f2*s2;
    s1 = signal[1];
    res[1] = f0*s1 + f1*s2 + f2*s1;
    s2 = signal[2];
    res[2] = f0*s2 + f1*s0 + f2*s2;
} while (…);
```

Expose Independent Operations

- Hide instruction latency
- Use local variables to expose independent operations that can execute in parallel or in a pipelined fashion
- Balance the instruction mix (what functional units are available?)

```java
f1 = f5 * f9;
f2 = f6 + f10;
f3 = f7 * f11;
f4 = f8 + f12;
```

Copy Optimization

- Copy input operands or blocks
  - Reduce cache conflicts
  - Constant array offsets for fixed size blocks
  - Expose page-level locality

Recursive Data Layouts

- Copy optimization often works because it improves spatial locality
- A related (recent) idea is to use a recursive structure for the matrix
- There are several possible recursive decompositions depending on the order of the sub-blocks
- This figure shows Z-Morton Ordering
- See papers on “cache oblivious algorithms” and “recursive layouts”

Advantages:
  - the recursive layout works well for any cache size
Disadvantages:
  - The index calculations to find A[i,j] are expensive
  - Implementations switch to column-major for small sizes

Strassen’s Matrix Multiply

- The traditional algorithm (with or without tiling) has O(n^3) flops
- Strassen discovered an algorithm with asymptotically lower flops
  - O(n^2.81)
- Consider a 2x2 matrix multiply, normally takes 8 multiplies, 7 adds
- Strassen does it with 7 multiplies and 18 adds

```
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
</tr>
</tbody>
</table>
```

Strassen (continued)

```
T(n) = Cost of multiplying nxn matrices
      = 7*T(n/2) + 18*(n/2)^2
      = O(n log_2 7)
      = O(n^2.81)
```

- Asymptotically faster
- Several times faster for large n in practice
- Cross-over depends on machine
- Available in several libraries
- Caveats
  - Needs more memory than standard algorithm
  - Can be less accurate because of roundoff error
  - Current world’s record is O(n 2.376… )
- Why does Hong/Kung theorem not apply?
Locality in Other Algorithms

- The performance of any algorithm is limited by $q$.
- In matrix multiply, we increase $q$ by changing computation order.
  - Increased temporal locality.
- For other algorithms and data structures, even hand-transformations are still an open problem.
  - Sparse matrices (reordering, blocking).
  - Trees (B-Trees are for the disk level of the hierarchy).
  - Linked lists (some work done here).

Summary

- Performance programming on uniprocessors requires understanding of memory system.
- Understanding of fine-grained parallelism in processor.
- Simple performance models can aid in understanding.
  - Two ratios are key to efficiency (relative to peak):
    1. Computational intensity of the algorithm:
      - $q = \frac{f}{m} = \frac{\# \text{ floating point ops}}{\# \text{ slow memory ops}}$
    2. Machine balance in the memory system:
      - $t_{m}/t_f = \text{time for slow memory operation} / \text{time for floating point operation}$
- Blocking (tiling) is a basic approach.
- Techniques apply generally, but the details (e.g., block size) are architecture-dependent.
- Similar techniques are possible on other data structures and algorithms.
- Now it's your turn: Homework 1
  - Work in teams of 2 or 3 (assigned this time).

Reading for Today

- Sourcebook Chapter 3, note that chapters 2 and 3 cover the material of lecture 2 and lecture 3, but not in the same order.
- Web pages for reference:
  - BeBOP Homepage
  - ATLAS Homepage
  - BLAS (Basic Linear Algebra Subroutines), Reference for (unoptimized) implementations of the BLAS, with documentation.
  - LAPACK (Linear Algebra PACKage), a standard linear algebra library optimized to use the BLAS effectively on unprocessors and shared memory machines (software, documentation and reports).
  - ScalAPACK (Scalable LAPACK), a parallel version of LAPACK for distributed memory machines (software, documentation and reports).
- Tuning Strassen’s Matrix Multiplication for Memory Efficiency

Questions You Should Be Able to Answer

1. What is the key to understand algorithm efficiency in our simple memory model?
2. What is the key to understand machine efficiency in our simple memory model?
3. What is tiling?
4. Why does block matrix multiply reduce the number of memory references?
5. What are the BLAS?
6. What is LAPACK? ScalAPACK?
7. Why does loop unrolling improve uniprocessor performance?

Reading Assignment

- Next week: Current high performance architectures
  - Cray X1
  - Blue Gene L
  - Clusters
    - http://www.mirror.ac.uk/sites/www.beowulf.org/papers/ICPP95/
- Optional
  - Chapter 1.2 of the “Sourcebook of Parallel Computing”
  - Alternative to Beowulf paper:
    - http://now.cs.berkeley.edu/Case/case.html

Review from Last Lecture
Membench: What to Expect

- Consider the average cost per load
  - Plot one line for each array size, time vs. stride
  - Small stride is best: if cache line holds 4 words, at most ¼ miss
  - If array is smaller than a given cache, all those accesses will hit (after the first run, which is negligible for large enough runs)
  - Picture assumes only one level of cache
  - Values have gotten more difficult to measure on modern procs

Memory Hierarchy on a Sun Ultra-2i

- L2: 2 MB, 12 cycles (36 ns)
- L1: 16 KB, 2 cycles (6 ns)
- Mem: 396 ns (132 cycles)
- Sun Ultra-2i, 333 MHz
- L1: 16 B line
- L2: 64 byte line
- 8 K pages
- See www.cs.berkeley.edu/~yelick/arvindk/t3d-isca95.ps for details

PHiPAC: Portable High Performance ANSI C

- Speed of n-by-n matrix multiply on Sun Ultra-1/170, peak = 330 MFlops