VECTOR COMPUTING

Scalar

X

* 

Y

X * Y

SIMD

X3  X2  X1  X0

* 

Y3  Y2  Y1  Y0

X3 * Y3  X2 * Y2  X1 * Y1  X0 * Y0

Claude TADONKI
Mines ParisTech – Paris/France

Seminar at Universidad Santiago de Chile
August 6, 2014

SANTIAGO - CHILE
Could you explain how to get the 204.8 GFLOPS for BlueGene/Q?

Demo: cygwin /home/codeSIMD ./docompil ./go

Could you explain how to get the 204.8 GFLOPS for BlueGene/Q?
A **SIMD** machine simultaneously operates on tuples of atomic data (*one instruction*).

**SIMD** is opposed to **SCALAR** (the traditional mechanism).

**SIMD** is about exploiting parallelism in the data stream (*DLP*), while superscalar **SISD** is about exploiting parallelism in the instruction stream (*ILP*).

**SIMD** is usually referred as **VECTOR COMPUTING**, since its basic unit is the **vector**.

Vectors are represented in what is called **packed data format** stored into **vector registers**.

On a given machine, the length of the vector registers and their number are fixed and determine the hardware **SIMD** potential.

**SIMD** can be implemented on using specific extensions **MMX, SSE, AVX, …**
Pipeline Floating Point Computation (multi-stage)

Consider the 6 steps (stages) involved in a floating-point addition on a sequential machine with IEEE arithmetic hardware:

A. exponents are compared for the smallest magnitude.
B. exponents are equalized by shifting the significand smaller.
C. the significands are added.
D. the result of the addition is normalized.
E. checks are made for floating-point exceptions such as overflow.
F. rounding is performed.

A scalar implementation of adding two array of length n will require 6n steps
A pipeline implementation of adding two array of length n will require 6 + (n-1) steps
Some architectures provide a wider overlapping by chaining the pipelines.
Roughly speaking, a p-length vector computation on a given n-array needs n/p steps.
Depending on the architecture, pipeline processing applies to # operations (arith, logical).
Pipeline feature is usually covered in the topic of Instruction Level Parallelism (ILP)

Could you identify and explain other type of pipeline in a standard computation scheme?

FP comp // Integer comp // Load // store // ….
SIMD Implementation

**SIMD: Continuous Evolution**

<table>
<thead>
<tr>
<th>Year</th>
<th>Version</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>SSE</td>
<td>70 instr. Single-Precision Vectors Streaming operations</td>
</tr>
<tr>
<td>2000</td>
<td>SSE2</td>
<td>144 instr Double-precision Vectors 8/16/32 64/128-bit vector integer</td>
</tr>
<tr>
<td>2004</td>
<td>SSE3</td>
<td>13 instr Complex Data</td>
</tr>
<tr>
<td>2006</td>
<td>SSSE3</td>
<td>32 instr Decode</td>
</tr>
<tr>
<td>2007</td>
<td>SSE4.1</td>
<td>47 instr Video Graphics building blocks Advanced vector instr</td>
</tr>
<tr>
<td>2008</td>
<td>SSE4.2</td>
<td>8 instr String/XML processing POP-Count CRC</td>
</tr>
<tr>
<td>2009</td>
<td>AES-NI</td>
<td>7 instr Encryption and Decryption Key Generation</td>
</tr>
<tr>
<td>2010\11</td>
<td>AVX</td>
<td>~100 new instr. ~300 legacy SSE instr updated 256-bit vector 3 and 4-operand instructions</td>
</tr>
</tbody>
</table>
**SIMD Implementation**

**MMX™**
- Vector size: 64bit
- Data types: 8, 16 and 32 bit integers
- VL: 2, 4, 8
- For sample on the left: Xi, Yi 16 bit integers

**Intel® SSE**
- Vector size: 128bit
- Data types: 8, 16, 32, 64 bit integers
- VL: 2, 4, 8, 16
- Sample: Xi, Yi bit 32 int / float

**Intel® AVX**
- Vector size: 256bit
- Data types: 32 and 64 bit floats
- VL: 4, 8, 16
- Sample: Xi, Yi 32 bit int or float

**Intel® MIC**
- Vector size: 512bit
- Data types: 32 and 64 bit integers, 32 and 64bit floats (some support for 16 bits floats)
- VL: 8, 16
- Sample: 32 bit float

**MMX = MultiMedia eXtension**
**SSE = Streaming SIMD Extension**
**AVX = Advanced Vector Extensions**
**MIC = Many Integrated Core**
SSE (Overview)

- **SSE = Streaming SIMD Extensions**

- SEE programming can be done either through *(inline) assembly* or from a high-level language (C and C++) using *intrinsics*.

- The `{x,e,p}mmintrin.h` header file contains the declarations for the SSEx instructions intrinsics.
  - `xmmintrin.h` -> SSE
  - `emmintrin.h` -> SSE2
  - `pmmintrin.h` -> SSE3

- SSE instruction sets can be enabled or disabled. If disable, SSE instructions will not be possible. It is recommended to leave this BIOS feature enabled by default. In any case MMX (MultiMedia eXtensions) will still available.

- compile your SSE code with "gcc -o vector vector.c -msse -msse2 -msse3"

- SSE intrinsics use types `__m128` *(float)*, `__m128i` *(int, short, char)*, and `__m128d` *(double)*

- Variable of type `__m128`, `__m128i`, and `__m128d` (exclusive use) maps to the XMM[0-7] registers *(128 bits)*, and automatically aligned on 16-byte boundaries.

- Vector registers are `xmm0, xmm1, …, xmm7`. Initially, they could only be used for *single precision* computation. Since SSE2, they can be used for *any primitive data type*. 
Vector variables can be connected to scalar variables (arrays) using one of the following ways:

```c
float a[N] __attribute__((aligned(16)));
__m128 *ptr = (__m128*)a;

prt[i] or *(ptr+i) represents the vector {a[4i], a[4i+1], a[4i+2], a[4i+3]}
```

```c
float a[N] __attribute__((aligned(16)));
__m128 mm_a;
mm_a = _mm_load_pd(&a[i]); // here we explicitly load data into the vector

mm_a represents the vector {a[4i], a[4i+1], a[4i+2], a[4i+3]}
```

Using the above connection, we can now use SSE instruction to process our data. This can be done through:

- *(inline) assembly
- intrinsics (interface to keep using high-level instructions to perform vector operations)

Pros and cons of using (inline)assembly versus intrinsics.
SSE (basic assembly instructions)

Data Movement Instructions

- **MOVUPS** - Move 128bits of data to an SIMD register from memory or SIMD register. Unaligned.
- **MOVAPS** - Move 128bits of data to an SIMD register from memory or SIMD register. Aligned.
- **MOVHPS** - Move 64bits to upper bits of an SIMD register (high).
- **MOVLPS** - Move 64bits to lower bits of an SIMD register (low).
- **MOVHLPS** - Move upper 64bits of source register to the lower 64bits of destination register.
- **MOVLHPS** - Move lower 64bits of source register to the upper 64bits of destination register.
- **MOVMSKPS** - Move sign bits of each of the 4 packed scalars to an x86 integer register.
- **MOVSS** - Move 32bits to an SIMD register from memory or SIMD register.

Arithmetic Instructions

- **ADDPS** - Adds operands
- **SUBPS** - Subtracts operands
- **MULPS** - Multiplies operands
- **DIVPS** - Divides operands
- **SQRTPS** - Square root of operand
- **MAXPS** - Maximum of operands
- **MINPS** - Minimum of operands
- **RCPPS** - Reciprocal of operand
- **RSQRTPS** - Reciprocal of square root of operand

Logical Instructions

- **ANDPS** - Bitwise AND of operands
- **ANDNPS** - Bitwise AND NOT of operands
- **ORPS** - Bitwise OR of operands
- **XORPS** - Bitwise XOR of operands
Shuffling offers a way to
- change the order of the elements within a single vector or
- combine the elements of two separate registers.

**Shuffle Instructions**
- **SHUFPS** - Shuffle numbers from one operand to another or itself.
- **UNPCKHPS** - Unpack high order numbers to an SIMD register.
- **UNPCKLPS** - Unpack low order numbers to a SIMD register.

The **SHUFPS** instruction takes two SSE registers and an 8 bit hex value. (elements are numbered from right to left !!!)

- The **first two** elements of the destination operand are overwritten by any two elements of the destination register.
- The **third and fourth** elements of the destination register are overwritten by any two elements from the source register.
- The **hex string** is used to tell the instruction which elements to shuffle.
  - 00, 01, 10, and 11 are used to access elements within the registers.

**Examples**
- **SHUFPS XMM0, XMM0, 0x1B**  // 0x1B = 00 01 10 11 and reverses the order of the elements
- **SHUFPS XMM0, XMM0, 0xAA**  // 0xAA = 10 10 10 10 and sets all elements to the 3rd element

Write the shuffling instruction to obtain (a2, a3, a0, a1) from (a3, a2, a1, a0) in XMM0
What is XMM0 after **SHUFPS XMM0, XMM0, 93h**?
What is XMM0 after **SHUFPS XMM0, XMM0, 39h**?
SSE (assembly examples)

```c
// Use sse to multiply vector elements by a real number a * b
vector4 sse_vector4_multiply(const vector4 &op_a, const float &op_b)
{
    vector4 ret_vector;
    __m128 f = _mm_set1_ps(op_b); // Set all 4 elements to op_b
    __asm
    {
        MOV    EAX, op_a       // Load pointer into CPU reg
        MOVUPS XMM0, [EAX]    // Move the vectors to SSE regs
        MULPS  XMM0, f         // Multiply elements
        MOVUPS [ret_vector], XMM0 // Save the return vector
    }
    return ret_vector;
}
```

```c
// Use sse to add the elements of two vectors a + b
vector4 sse_vector4_add(const vector4 &op_a, const vector4 &op_b)
{
    vector4 ret_vector;
    __asm
    {
        MOV    EAX, op_a       // Load pointers into CPU regs
        MOV    EBX, op_b
        MOVUPS XMM0, [EAX]    // Move the vectors to SSE regs
        MOVUPS XMM1, [EBX]
        ADDPS  XMM0, XMM1      // Add elements
        MOVUPS [ret_vector], XMM0 // Save the return vector
    }
    return ret_vector;
}
```
SSE (assembly examples)

We need to write a SSE code to calculate the cross product

R.x = A.y * B.z - A.z * B.y
R.y = A.z * B.x - A.x * B.z
R.z = A.x * B.y - A.y * B.x

Complete the following code

```c
// Use sse to add the elements of two vectors a + b
vector4 sse_vector4_cross_product(const vector4 &op_a, const vector4 &op_b){
    vector4 ret_vector;
    __asm
    {
      MOV      EAX, op_a          // Load pointers into CPU regs
      MOV      EBX, op_b
      MOVUPS   XMM0, [EAX]        // Move the vectors to SSE regs
      MOVUPS   XMM1, [EBX]

      MOVUPS   [ret_vector], XMM0 // Save the return vector
    }
    return ret_vector;
}
```
We need to write a SSE code to calculate the cross product

\[
\begin{align*}
R.x &= A.y \times B.z - A.z \times B.y \\
R.y &= A.z \times B.x - A.x \times B.z \\
R.z &= A.x \times B.y - A.y \times B.x
\end{align*}
\]

Complete the following code

```cpp
// Use sse to add the elements of two vectors a + b
vector4 sse_vector4_cross_product(const vector4 &op_a, const vector4 &op_b){
    vector4 ret_vector;
    __asm {
        MOV    EAX, op_a              // Load pointers into CPU regs
        MOV    EBX, op_b
        MOVUPS XMM0, [EAX]           // Move the vectors to SSE regs
        MOVUPS XMM1, [EBX]
        MOVAPS XMM2, XMM0
        MOVAPS XMM3, XMM1
        SHUFPS XMM0, XMM0, 0xD8
        SHUFPS XMM1, XMM1, 0xE1
        MULPS  XMM0, XMM1
        SHUFPS XMM2, XMM2, 0xE1
        SHUFPS XMM3, XMM3, 0xD8
        MULPS  XMM2, XMM3
        SUBPS  XMM0, XMM2
        MOVUPS [ret_vector], XMM0   // Save the return vector
    }
    return ret_vector;
}
```
SSE (common intrinsics)

```c
_mm_add_ps(__m128 a, __m128 b)
_mm_sub_ps(__m128 a, __m128 b)
_mm_mul_ps(__m128 a, __m128 b)
_mm_div_ps(__m128 a, __m128 b)
_mm_sqrt_ps(__m128 a, __m128 b)
_mm_min_ps(__m128 a, __m128 b)
_mm_max_ps(__m128 a, __m128 b)

_mm_cmpeq_ps(__m128 a, __m128 b)
_mm_cmplt_ps(__m128 a, __m128 b)
_mm_cmpgt_ps(__m128 a, __m128 b)

_mm_and_ps(__m128 a, __m128 b)
__mm_prefetch(__m128 a, _MM_HINT_T0)
```

Pros and cons of the prefetch.
void scalar_sqrt(float *a) {
    int i;
    for (i = 0; i < N; i++)
        a[i] = sqrt(a[i]);
}

void sse_sqrt(float *a) {
    // We assume N % 4 == 0.
    int nb_iters = N / 4;
    __m128 *ptr = (__m128*)a;
    int i;
    for (i = 0; i < nb_iters; i++, ptr++, a += 4)
        _mm_store_ps(a, _mm_sqrt_ps(*ptr));
}

Tadonki@TADONKI-PC ~/vector
$ ./test
Running time of the scalar code: 0.286017
Running time of the SSE code: 0.031001

10 times faster !!!!!!!
Write the SSE loop equivalent to the following scalar loop (use vectors mm_d, mm_a, mm_b, mm_c).

for(i = 0; i < N; i++)
    d[i] = (a[i] - b[i])*c[i];

for( i = 0; i < N; i+= 4){
    mm_a = _mm_load_ps(&a[i]);
    mm_b = _mm_load_ps(&b[i]);
    mm_c = _mm_load_ps(&c[i]);
    mm_r = _mm_add_ps( mm_a, mm_b );
    mm_a = _mm_mul_ps( mm_r , mm_c );
    _mm_store_ps( &r[i], mm_a );
}

Write the SSE loop equivalent to the following scalar loop (use vectors mm_c, mm_a, mm_b).

for(i = 0; i < N; i+= 2){
    c[2*i]   = (a[2*i] - b[2*i+1]);
    c[2*i+1] = (a[2*i+1] - b[2*i]);
}

Write the SSE loop equivalent to the following scalar loop (typedef struct {float re; float im} complex:).

for(i = 0; i < N; i++) c[i] = multiply(a[i], b[i]);

Write the SSE loop equivalent to the following scalar loop

for(i = 0; i < N; i++) b[i] = 2*a[i] + 1;

Back to the cross product