Measuring the Performance of Multimedia Instruction Sets

Nathan Slingerland and Alan Jay Smith
{slingn, smith}@cs.berkeley.edu

Computer Science Division
EECS Department
University of California at Berkeley

December 22, 2000

Abstract

Many microprocessor instruction sets include instructions for accelerating multimedia applications such as DVD playback, speech recognition and 3D graphics. Despite general agreement on the need to support this emerging workload, there are considerable differences between the instruction sets that have been designed to do so. In this paper we study the performance of five instruction sets on kernels extracted from a broad multimedia workload. Each kernel was recoded in the assembly language of the five multimedia extensions. We compare the performance of each extension against other architectures as well as to the original compiled C performance. From our analysis we determine how well multimedia workloads map to current architectures, what was useful and what was not. We also propose two enhancements to current architectures: strided memory operations, and superwide registers.

1 Introduction

Specialized instructions have been introduced by microprocessor vendors to support the specialized computational demands of multimedia applications. The mismatch between wide data paths and the relatively short data types found in multimedia applications has lead the industry to embrace SIMD (single instruction, multiple data) style processing. Unlike traditional forms of SIMD computing in which multiple individual processors execute the same instruction, multimedia instructions are executed by a single processor, and pack multiple short data elements into a single wide (64 or 128-bit) register, with all of the subelements being operated on in parallel.

The goal of this paper is to quantify how architectural differences between multimedia instruction sets translate into differences in performance. Prior studies have primarily focused on a single instruction set in isolation and have measured the performance on sets of kernels taken from vendor provided libraries [Nguy99], [Bhar98], [Allen99], [Chen96], [Rice96], [Naka96], [Rang99]. Our contribution is unique as we do not focus exclusively on a single architecture, and we study the performance of kernels derived from a real measured workload rather than those that have been established a priori. Our results are obtained from actual hardware measurements rather than through simulation, instilling confidence in our results.

Section 2 summarizes the multimedia workload we studied, and details the sixteen computationally important kernels which we extracted from it. Our methodology for recoding the kernels with multimedia instructions and their measurement is described in Section 3. An overview of the five instructions sets, and their implementations, is given in Section 4.

Our analysis is divided into two parts. Section 5 reflects our experience in coding the kernels, and gives insight into the useful and less than useful features of the multimedia instruction sets studied. In Section 6 we compare the performance of the five instruction sets both against one another, as well as their relative improvement over compiled (optimized) C code. Finally, in Section 7 we propose two new directions for multimedia architectures on general purpose microprocessors: strided memory operations and superwide registers.

2 Workload

The lack of a standardized multimedia benchmark has meant that workload selection is the most difficult aspect of any study of multimedia. It was for this reason that we developed the Berkeley multimedia workload [Sling00a]. In selecting the component applications we strove to cover as many types of media processing as possible: image compression (DjVu, JPEG), 3D graphics (Mesa, POVray), document rendering (Ghostscript), audio synthesis (Timidity), audio compression (ADPCM, LAME, mpg123), video compression (MPEG-2 at DVD and HDTV resolutions), speech synthesis (Rsynth), speech compression (GSM), speech recognition (Rasta) and video game (Doom) applications. Open source software was used both for its portability (allowing for cross platform comparisons) and the fact that we could analyze the source code

Funding for this research has been provided by the State of California under the MICRO program, and by Cisco Corporation, Fujitsu Microelectronics, IBM, Intel Corporation, Maxtor Corporation, Microsoft Corporation, Sun Microsystems, Toshiba Corporation and Veritas Software Corporation.

directly.

To evaluate the various multimedia instruction sets, we hand coded kernels selected from the elements of the Berkeley multimedia workload. Those kernels were chosen based on their computational significance and their suitability for SIMD optimization. Table 1 lists the kernel codes examined. Both Mesa kernels appear to take up a relatively small amount of application CPU time, as software rendering (computing all stages of a rendering pipeline) was used in the portable version of these applications. It was for this reason that rasterization kernels were not included in the kernels studied due to the ubiquity of 3D accelerator cards which offload this from the CPU, and will continue to do so in the foreseeable future. Although we expect that when enough CPU cycles become available, much of the 3D rendering workload will be moved back onto the CPU and done in software (for cost savings), the current trend is moving in the opposite direction. First generation 3D accelerator cards took care of the rasterization stage, but not 3D geometry computations. Current 3D accelerator cards have also taken on the burden of geometry computations, indicating that the growth in complexity of 3D environments is outpacing that of CPU performance, despite the best efforts of multimedia extensions.

3 Methodology

3.1 Berkeley Multimedia Kernel Library

Our goal is to measure the performance of existing multimedia instruction sets on our set of important multimedia kernels. Our first step was to we distill from our Berkeley multimedia workload [Sling00a] a set of computationally important kernel functions, from which we formed the *Berkeley multimedia kernel library* (BMKL). All of the parent applications in the Berkeley multimedia workload were modified to make calls to BMKL rather than internal functions. From a performance standpoint, no piece of code can be realistically extracted and studied in isolation. By measuring a piece of code from within a real system we can realistically see how the shared resources of a computer system, such as CPU, caches, TLB, memory, and registers, affect the code and are affected by it.

Encapsulating the kernels within a library with a well defined interface allowed for: 1) low overhead measurement code to be placed around library functions to make measurements as non-invasive as possible, 2) different versions of the library to be quickly substituted to aid testing, 3) the addition of new architectures to our study by starting with a copy of the C reference library and then implementing and debugging replacement SIMD assembly functions one at a time.

3.2 Coding Process

As with DSPs, the most efficient way to program with multimedia extensions is to have an expert programmer tune software using assembly language [Kuro98]. Although this is more tedious and error prone than other methods such as hand coded standard shared libraries or automatic code generation by a compiler, it is a method that is available on every platform, and allows for great flexibility and precision when coding. We chose to code each kernel in assembly language ourselves rather than measure vendor supplied optimized libraries to prevent differences in programmer ability and time spent coding between the vendors' from potentially skewing our comparison.

All of the assembly codes in our study were written by the same programmer (Slingerland) with an approximately equal amount of time spent on each platform. Programming tools for cycle-accurate instruction scheduling through simulation were not used in order to equalize differences between the tools available on each platform. Instructions were scheduled sanely by keeping data consumption as far from data use as possible, unrolling loops when of performance benefit, and manually scheduling instructions so as to take advantage of multiple functional units.

For reasons of practicability, we limited our optimizations to the kernel level; we did not rewrite entire applications. This had ramifications for data alignment and data structure layout on some architectures. In some cases it was not practical to code a SIMD version of a kernel if an instruction set lacked the requisite functionality. For example, Sun's VIS and DEC's MVI do not support partitioned floating point, so floating point kernels were not recoded for these platforms. DEC's MVI also does not contain any data communication (e.g. permute, mix, merge) or partitioned integer multiply instructions. If there was no compelling opportunity for performance gain, kernels were not recoded from C. The C version was considered to be a particular platform's "solution" to a kernel when the needed SIMD operations were not provided. It might be supposed that hand coding would be superior to compiler generated code anyway, even without SIMD instructions. Although this may have been true at one time when instruction set architectures were designed with assembly language programmers in mind, modern instruction sets are targeted at compilers [Lawl92], [Patt96].

3.3 Measurement

Performance Monitoring Counters All of the microprocessors studied include performance monitoring counters to allow for interesting architectural events to be counted in real time during program execution. Although performance counters were sometimes used to guide our optimizations, their primary purpose was to be nearly cycle-accurate timers with which to measure the very short execution times of the kernels in the BMKL.

C Compilers and Optimization Flags The most architecturally tuned compiler on each architecture was used to compile the C reference version of the Berkeley Multimedia Kernel library. The optimization flags used were those which give the best general speedup and highest level of optimization without resorting to tuning the compiler flags to a particular kernel. The specific compiler and associated optimization flags used on each platform are listed given in [Sling00d].

Kernel Name	Source Application	Data Type	Sat Arith	Native Width	Src Lines	Satic Instr	%Static Instr	% CPU Cycles
Add Block	MPEG-2 Decode	8-bit (U)	√	64-bits	46	191	0.1%	13.7%
Block Match	MPEG-2 Encode	8-bit (U)		128-bits	52	294	0.4%	59.8%
Clip Test and Project	Mesa*	FP		limited	95	447	0.1%	0.8%
Color Space Conversion	JPEG Encode	8-bit (U)		limited	22	78	0.1%	9.8%
DCT	MPEG-2 Encode	16-bit (S)		128-bits	14	116	0.1%	12.3%
FFT	LAME*	FP		limited	208	981	4.4%	14.5%
Inverse DCT	MPEG-2 Decode	16-bit (S)	√	128-bits	75	649	0.3%	29.7%
Max Value	LAME*	32-bit (S)		unlimited	8	39	0.2%	12.0%
Mix	Timidity	16-bit (S)	√	unlimited	143	829	1.0%	35.7%
Quantize	LAME*	FP		unlimited	55	312	1.4%	15.3%
Short Term Analysis Filter	GSM Encode	16-bit (S)	√	128-bits	15	79	0.1%	20.2%
Short Term Synthesis Filter	GSM Decode	16-bit (S)	√	128-bits	15	114	0.2%	72.7%
Subsample Horizontal	MPEG-2 Encode	8-bit (U)	√	88-bits	35	244	0.3%	2.6%
Subsample Vertical	MPEG-2 Encode	8-bit (U)	√	unlimited	76	478	0.6%	2.1%
Synthesis Filtering	mpg123*	FP	√	512-bits	67	348	0.4%	39.6%
Transform and Normalize	Mesa*	FP		limited	51	354	0.1%	0.7%

Table 1: **Multimedia Kernels Studied** - from left to right, the colums list 1) primary data type specified as N-bit ({Unsigned,Signed}) integer or floating point (FP), 2) if saturating arithmetic is used, 3) native width; the longest width which does not load/store/compute excess unused elements, 4) static C source line count (for those lines which are executed), 5) percentage of total static instructions, 6) percentage of total CPU time spent in the kernel. The later three statistics are machine specific and are for the original C code on a Compaq DS20 (dual 500 MHz Alpha 21264, Tru64 Unix v5.0 Rev. 910) machine, compiled with GCC v2.8.1 (*) or DEC C v5.6-075.

4 Instruction Sets

In this paper we study five architectures with different multimedia instruction sets (Table 2 lists the parameters for the exact parts we used). Most of the instruction sets (AMD'S 3DNow!, DEC's MVI, Intel's MMX, Sun's VIS) use 64-bit wide registers, while Motorola's AltiVec and Intel's SSE are 128-bits wide. The size of the register file available on each architecture varied widely, ranging from 8 64-bit registers on the x86 based AMD Athlon and Intel Pentium III to 32 128-bit wide registers with Motorola's AltiVec extension.

All of the multimedia extensions support integer operations, although the types and widths of available operations vary greatly. The earliest multimedia instruction sets (e.g. Sun's VIS and DEC's MVI) had design goals limited by the immaturity of their target workload and unproven benefit to performance. Because of this, any approach which greatly modified the overall architecture or significantly affected die size was out of the question. Leveraging as much functionality as possible from existing chip architectures was a high priority. Thus, the Sun and DEC extensions do not implement partitioned floating point instructions. Witness also the difference between Intel's first multimedia extension, MMX (1997), which had as one of its primary design goals to not require any new operating system support, and Intel's SSE (1999) which added new operating system maintained state (8 128-bit wide registers) for the first time since the introduction of the Intel386 instruction set (1985).

The processors studied vary in terms of instruction latency as well as the throughput per cycle. Processor clock rates were all 500 MHz, with the exception of the Sun UltraSPARC IIi, for which only a 360 MHz system was available. Although multimedia extensions primarily focus on extracting data level parallelism, most modern microprocessors are also superscalar, and thereby allow for multiple multimedia instructions to be issued every cycle. All of the architectures we looked at are fully pipelined, so barring any data dependencies, one new SIMD instruction can begin per functional unit each clock cycle. Typically, there are separate functional units for SIMD integer and SIMD floating point processing, although on the x86 architectures they are combined. [Sling00c] surveys existing multimedia instruction sets in more detail.

Sun Sun's UltraSPARC IIi processor incorporates the VIS multimedia extension, which implements a set of SIMD integer instructions that share the existing UltraSPARC floating point register file. Partitioned multiplication is done through 8-bit multiplication primitive instructions. A graphics status register (GSR) is used to support data alignment and scaling for pack operations.

Intel Intel's Pentium III processor includes their original SIMD integer MMX extension, as well as the newer SIMD floating point SSE instruction set. MMX is a 64-bit wide SIMD integer extension, which is mapped onto the existing x87 floating point architecture and registers and introduces no new architectural state (registers or exceptions). SSE is a follow on to MMX which is primarily a SIMD floating point extension but also incorporates feedback on MMX from software vendors in the form of new integer instructions. Unlike MMX, the floating point side of SSE *does* add new architec-

	AMD Athlon	DEC Alpha 21264	Intel Pentium III	Motorola G4	Sun UltraSPARC IIi
Clock [Current] (MHz)	500 [1000]	500 [667]	500 [1000]	500 [500]	360 [480]
SIMD Extension(s)	MMX/3DNow!+	MVI	MMX/SSE	AltiVec	VIS
SIMD Instructions	57/24	13	57/70	162	121
First Shipped	June 1999	February 1998	February 1999	August 1999	November 1998
Transistors ($x10^6$)	22.0	15.2	9.5	6.5	5.8
Process (μm) [Die Size (mm^2)]	0.25 [184.0]	0.25 [225.0]	0.20 [104.6]	0.20 [83.0]	0.25 [147.5]
L1 $I Cache, $D Cache (Kbytes)	64, 64	64, 64	16, 16	32, 32	32, 64
L2 Cache (Mbytes)	0.5	4	0.5	1	2
Register File (# x width)	FP(8x64b)/FP(8x64b)	Int (31x64b)	FP(8x64b)/8x128b	32x128b	FP(32x64b)
Reorder Buffer Entries	72	35	40	6	12
Int, FP Multimedia Units	2 (combined)	2, 0	2 (combined)	3, 1	2, 0
Int Add Latency	2 [64b/cycle]	-	1 [64b/cycle]	1 [128b/cycle]	1 [64b/cycle]
Int Multiply Latency	3 [64b/cycle]	-	3 [64b/cycle]	3 [128b/cycle]	3 [64b/cycle]
FP Add Latency	4 [64b/cycle]	-	4 [64b/cycle]	4 [128b/cycle]	-
FP Multiply Latency	4 [64b/cycle]	-	5 [64b/cycle]	4 [128b/cycle]	-
FP ~1/sqrt Latency	3 [32b/cycle]	-	2 [64b/cycle]	4 [128b/cycle]	-

Table 2: **Microprocessors Studied** - All parameters are for the actual chips used in this study. Clock lists the speed for the specific part studied as well as the current (December 2000) maximum shipping clock speed. A SIMD register files may be shared with existing integer (Int) or floating point (FP) registers, or be separate. Note that some of the processors implement several multimedia extensions (e.g. Pentium III has MMX and SSE) - the corresponding parameters for each are separated with "/". A dash (-) indicates that a particular operation is not available on a given architecture, and so no latency and throughput numbers are given. DEC's MVI extension (on the 21264) does not include any of the listed operations, but all MVI instructions have a latency of 3 cycles [64b/cycle].[Burd] [Noer] [AMD99] [AMD00] [AMDWP] [Carl97] [Comp00] [Intel97] [Intel99a] [Kesh99] [Kohn95] [Moto00] [Norm98] [Sun97]

tural state to the Intel architecture with the addition of an 8 x 128-bit register file and exceptions to support IEEE compliant floating point operations. Although the SSE instruction set architecture and register file are defined to be 128-bits wide, the Pentium-III SSE execution units are actually 64-bits (two single precision floating point elements) wide in hardware. The instruction decoder translates 4-wide (128-bit) SSE instructions into pairs of 2-wide (64-bit) internal micro-ops.

AMD The AMD Athlon processor implements MMX (which was licensed from Intel), in addition to AMD's own 3DNow! extension which utilizes the same x87 floating point registers and basic instruction formats as MMX, but adds a partitioned single precision floating point data type. The Athlon processor actually extends 3DNow! with Enhanced 3DNow! that adds floating point and integer operations to make 3DNow! functionally equivalent to Intel's SSE extension.

DEC The DEC (now Compaq, but we will refer to it as DEC for historical consistency) Alpha 21264A processor includes the SIMD integer Motion Video Instructions (MVI) multimedia extension. It is the smallest of the multimedia instruction sets, weighing in with only 13 instructions. MVI shares the existing Alpha 32-register integer register file. Notably, no SIMD saturating addition/subtraction, multiplication, or shift instructions are included.

Motorola Motorola's MPC7400 (also known as the G4) processor utilizes their 128-bit wide SIMD AltiVec extension which supports a wide variety of integer data types, as well as partitioned single precision floating point. A dedicated 32 x 128-bit register file is implemented, along with four non-identical parallel, pipelined vector execution units. Hardware assisted software prefetching is implemented, where by a prefetch stream is set up by software, and fetched into the cache independently by hardware.

5 Analysis

In the next section we use our experience coding each of the sixteen kernels with five different multimedia extensions to determine: 1) existing architectural features that are useful, 2) features that have been implemented, but don't appear to be useful, and 3) significant bottlenecks in current multimedia architectures. Illustrating our discussion are code fragments both from the original C source code of each kernel algorithm, as well as the different SIMD implementations. The code fragments consist of a few of the key central lines of code from a given kernel. This gives an idea about the types of operations and data types used. The data types of all of the variables in our sample C code are specified in a platform independent way such that the prefix indicates the type: INT: signed integer, UINT: unsigned integer, FP: floating point, followed by N, the number of bits. The complete original C source code for each kernel as well as for the assembly language SIMD implementations of the BMKL are available

on the web at *http://www.cs.berkeley.edu/~slingn/research/.*

5.1 Register File and Data Path

Multimedia instruction sets can be broadly categorized according to the location and geometry of the register file upon which SIMD instructions operate. Alternatives include the reuse of the existing integer or floating point register files, or implementing an entirely separate one. The type of register file affects the width and therefore the number of packed elements that can be operated on simultaneously (vector length).

Integer Data Path Implementing multimedia instructions on the integer data path has the advantage that the functional units for shift and logical operations need not be replicated. Partitioned addition and subtraction are easily created by blocking the appropriate carry bits. Modifications to the integer data path to accommodate multimedia instructions can potentially adversely affect the critical path of the integer data-path pipeline [Kuro98]. On the x86 (AMD, Intel) and PowerPC (Motorola) architectures the integer data path is 32-bits wide, making a shared integer data path approach less compelling due to the limited amount of data level parallelism possible.

Floating Point Data Path The reuse of floating point rather than integer registers has the advantage of not being shared with pointers and loop and other control flow variables. In addition, multimedia and floating point instructions are not typically used simultaneously [Kuro98]. All of the architectures examined have floating point data paths which support at least double-precision (64-bit wide) operations, which for many architectures is wider than the integer data path.

Separate Data Path A separate data path has the advantage of simplifying pipeline control and increasing the overall number of registers. Disadvantages include the need for saving and restoring the new registers on context switches, as well as the relative difficulty and high overhead of moving data between register files.

SIMD Register Width Perhaps one of the most critical factors in SIMD instruction set design is deciding how long vectors will be. If an existing data path is to be reused, there is little choice, but when a new data path is to be designed it makes sense to ask how wide is wide enough. Too short of a vector length limits the ability to exploit data parallelism, while an excessively long vector length can degrade performance and increase the amount of clean up code overhead. The "native width" column of Table 1 specifies how each of the multimedia kernels fits into one of the following three categories:

1. unlimited width - Kernels that operate on data elements which are truly independent and are naturally arranged so as to be amenable to SIMD processing. The inner loops of these kernels can be strip mined at almost any width, with the increase in performance of a longer vector being directly proportional to the increase in vector length.
2. limited width - Although data elements are independent, there is overhead involved in rearranging input data so that it may be operated on in a SIMD manner with longer vectors. Thus, the performance advantage of longer vectors is limited by the overhead (which typically increases with vector length) required to employ them.
3. exact width - A kernel which has a precise natural width which can be considered to be the right match for the kernel. This width is the longest width which does not load, store, or compute excess unused elements.

Long vectors can be a problem when their length exceeds the natural width of an algorithm. A good example of this problem is the add block kernel, which operates on MPEG subblocks (8 x 8 arrays of pixels). One input array (bp) consists of signed 16-bit integer values and the other (rfp) of unsigned 8-bit integer (Algorithm 1).

Algorithm 1 Add Block - computed for each pixel in a subblock

```
INT16 *bp; UINT8 *rfp; INT32 tmp;
tmp = *bp++ + *rfp; /* Add */
*rfp = tmp>255 ?  255 :  (tmp<0 ?  0 :  tmp); /* Clip */
```

During the block reconstruction phase of motion compensation in the decoder, a block of pixels is reconstituted by summing the pixels in different subblocks. Consider the AltiVec implementation of the add block kernel (Algorithm 2). Motorola's AltiVec is the only SIMD integer extension examined which is 128-bits wide; Intel's SSE is only 128-bits wide for packed floating point operations. Each time a row of the 8x8 rfp subblock is loaded on a 128-bit wide architecture, one half of the vector is useless data which will be thrown away when the rfp values are expanded to 16-bits.

Algorithm 2 AltiVec Add Block Fragment

```
;; unaligned vector load
lvx     v3, 0, r3         ; v3: vector MSQ for initial bp0..bp7 vector
lvx     v4, r11, r3       ; v4: vector LSQ
;; unaligned vector load
lvx     v0, 0, r4         ; v0: vector MSQ
lvx     v1, r11, r4       ; v1: vector LSQ
lvsl    v2, 0, r4         ; v2: vector alignment mask for vperm
vxor    v10, v10, v10     ; v10: 0
vperm   v0, v0, v1, v2    ; v0: rfp:|0|1|2|3|4|5|6|7|X|X|X|X|X|X|X|X|
vperm   v3, v3, v4, v5    ; v3: | bp0 | bp1 | bp2 | bp3 | bp4 | bp5 |
bp6 | bp7 |
addi    r3, r3, 16        ; r3: bp += 8 (pointer to INT16)
vmrghb  v1, v10, v0       ; v0: | rfp0| rfp1| rfp2| rfp3| rfp4| rfp5|
rfp6| rfp7|
vaddshs v1, v3, v1        ; v1: bp + rfp [0..7]
vpkshus v1, v1, v1        ; v1: rfp:|0|1|2|3|4|5|6|7|0|1|2|3|4|5|6|7|
stvewx  v1, 0, r4         ; store rfp [0..3]
stvewx  v1, r12, r4       ; store rfp [4..7]
add     r4, r4, r5        ; r4: rfp += (iincr + 8) (pointer to UINT8)
vmov    v3, v4            ; move current LSQ to next MSQ
```

As we saw in Table 1, most multimedia kernels are either unlimited/limited or have most often have an exact required width of 128-bits. The remaining exact native widths (64-, 88- and 512-bits) came up only once each. Thus, we consider a total vector length of 128-bits to be best.

Number of Registers Multimedia applications (and their kernels) can generally take advantage of quite large register files. Not coincidentally, MicroUnity's dedicated media processor chip has a 64 x 64-bit register file, which can also be accessed as 128 x 32-bits [Hans96], while the Philips Trimedia TM-1 has a 128 x 32-bit register file [Rath96].

As an example of where large numbers of registers are useful in our workload, consider the DCT and IDCT kernels (fragments of the original codes are given in Algorithms 3 and 4). The discrete cosine transform (DCT) is the algorithmic centerpiece to many lossy image compression methods. It is similar to the discrete Fourier transform (DFT) in that it maps values from the time domain to the frequency domain, producing an array of coefficients representing frequencies [Kien99]. The inverse DCT maps in the opposite direction, from the frequency to the time domain.

Algorithm 3 DCT

```
extern FLOAT64 c[8][8]; /* transform coefficients */
INT16 block[8][8]; FLOAT64 sum; FLOAT64 tmp[8][8];
  for (INT32 i=0; i<8; i++)
    for (int j=0; j<8; j++) {
      sum = 0.0;
      for (int k=0; k<8; k++)
        sum += c[j][k] * block[i][k];
      tmp[i][j] = sum;
    }
```

Algorithm 4 Inverse DCT - only row computation shown

```
INT32 x0,x1,x2,x3,x4,x5,x6,x7,x8
  x7 = x8 + x3;
  x8 -= x3;
  x3 = x0 + x2;
  x0 -= x2;
  x2 = (181*(x4+x5)+128)>>8;
  x4 = (181*(x4-x5)+128)>>8;
```

A 2D DCT or IDCT is efficiently computed as 1D transforms on each row followed by 1D transforms on each column and then scaling appropriately. A SIMD approach requires that multiple data elements from several iterations be operated on in parallel for the greatest efficiency. This is straightforward for the 1D column DCT, since the corresponding elements of each loop iteration are adjacent in memory (assuming a row-major storage format). A 1D row DCT is more problematic since the corresponding elements of adjacent rows are not. A matrix transposition (making corresponding "row" elements adjacent in memory), then performing the desired computation, and transposing the matrix back again (to put the resulting data back in the correct configuration) can be an effective way to compute the 1D row step of a 2D transform. However, this was only of performance benefit for those architectures whose register files were able to hold the entire 16-bit 8x8 matrix at once. Since the DCT and IDCT both operate on 8x8 2D matrices of 16-bit signed values, they require at least 16 64-bit registers, or 8 128-bit registers.

5.2 Data Types

The data types supported by the different multimedia instruction sets include {signed, unsigned}{8, 16, 32, 64} bit values, as well as single precision floating point. Most of the instruction sets do not support all of these types, and usually only a subset of operations on each. To determine which data types and operations are useful, we broke down the dynamic SIMD instruction counts on each architecture in two ways: 1) data type distribution per instruction class (e.g. add, multiply) and 2) data type distribution per kernel. Tables of these categorizations are available in [Sling00d].

In general, the video and imaging kernels (add block, block match, color space, DCT, IDCT, subsample horizontal, subsample vertical) utilize 8- and 16-bit operations. Audio kernels (FFT, max val, mix stereo, quantize, short term analysis filtering, short term synthesis filtering, synthesis filtering) either use 16-bit values or floating point, while the 3D kernels (clip test, transform) are limited almost exclusively to floating point.

Integer Although image and video data is typically stored as packed unsigned 8-bit values, intermediate processing usually requires precision greater than 8-bits. Other than for width promotion, most 8-bit functionality is wasted on our set of multimedia kernels. In general, *storage data types* (how data is stored in memory or on disk), although narrow and therefore potentially offering the greatest degrees of parallelism, are simply too narrow for intermediate computations to occur without overflow. A few operations inherently produce results that can not overflow the input data type. For example, although an N-bit average operation internally utilizes $N+1$ bits of precision to sum its two operands, the result is rounded and shifted back to N-bits before being stored to a register. Other operations such as the sum of absolute differences (SAD) produce a scalar result which fits in a destination register of the same width as the packed operands or a scalar register.

The signed 16-bit data type is the most heavily used because it is both the native data type for audio and speech data, as well as the typical intermediate data type for video and imaging. On the wide end of the spectrum, 32-bit and longer data types are typically only used for accumulation and simple data communication operations such as alignment. Operations tend to be limited to addition and subtraction (for accumulation), width promotion and demotion (for converting to a narrower output data type) and shifts (for data alignment).

Floating Point Single precision floating point plays an important role in many of the multimedia kernels such as the geometry stage of 3D rendering (the clipping and transform kernels) and the FFT, where the wide dynamic range of floating point is required. Only Intel's recently announced SSE2 extension will offer a packed double precision data type, to be targeted at applications other than multimedia such as scientific and engineering workloads, as well as advanced 3D geometry such as is used in raytracing [Intel00a], [Intel00b].

5.3 Operations

One of our primary goals is to separate useful SIMD operations from those that are not. The large differences in current multimedia instruction sets for general purpose processors are fertile ground for making such a determination because many different design choices have been made. In [Sling00d] we provide a table of SIMD instruction set functionality broken down per kernel, the important points of which we discuss here. Our analysis assumes that SIMD extensions are targeted solely at the domain of multimedia applications. In some cases, the targeted applications during the design of a multimedia extension included DSP applications and others which are not reflected in the Berkeley multimedia workload.

5.3.1 Arithmetic

Modulo/Saturation *Modulo arithmetic* "wraps around" to the next representable value when overflow occurs, while *saturating arithmetic* clamps the output value to the highest or lowest representable value for positive and negative overflow respectively. Saturating arithmetic is useful both because of its desirable aesthetic result in pixel based computations (video and imaging) as well as the fact that it allows for overflow in multiple packed elements to be dealt with efficiently. When adding pixels, modulo addition is undesirable since if overflow occurs a small change in operand values may result in a glaring visual difference (e.g. adding two white pixels results in a black pixel). If overflow within packed elements were to be handled similar to traditional scalar arithmetic, an overflowing SIMD operation would have to be repeated and tested serially to determine which element overflowed. The added cost of saturating arithmetic is that unlike modulo operations, for which the same instruction works for both unsigned and signed (2's complement) values, saturating arithmetic necessarily requires separate instructions since the values must be interpreted by the hardware as a particular data type.

Modulo computations are important because they allow for the results of SIMD optimized codes to be numerically identical to existing scalar algorithms. This is sometimes an important consideration for the sake of compatibility and comparability. The kernels in the BMKL which employ saturating arithmetic are noted in Table 1. From this it is clear that the most important types for saturating arithmetic are unsigned 8-bit and signed 16-bit integers. The IDCT kernel clamps to a signed 9-bit range [-256..+255], which can be accomplished through a pair of max/min operations; we discuss these in more detail later. Saturating 32-bit operations are of little value since overflow is usually not a concern for such a wide data type.

Shift SIMD shift operations are extremely important for supporting fixed point integer arithmetic. A common sequence of operations is the multiplication of an N-bit integer by an M-bit fixed point fractional constant, producing an $(N+M)$-bit result, with a binary point at the M^{th} most significant bit. At the end of computation, the final result is rounded by adding the fixed point fraction representing $\frac{1}{2}$, and then shifting the sum right M-bits to eliminate the fractional bits. Shifts are important operations for all data types, and are critical for fixed point integer arithmetic, as well as providing an inexpensive way to perform multiplication and division by powers of two.

Min/Max Min and max output the minimum or maximum values of the corresponding elements in two partitioned input registers, respectively. A max instruction is clearly useful in the maximum value search kernel, which searches through an array of signed 32-bit integers for the greatest maximum absolute value in the array. Max and min instructions have other less obvious uses as well. Signed minimum and maximum operations are often used with a constant second operand to saturate results to arbitrary ranges. The IDCT kernel clips its output value range to -256...+255 (9-bit signed integer), which does not correspond to the data types supported by any of the multimedia extensions. Algorithm 5 demonstrates clamping to arbitrary boundaries for the Intel implementation of the IDCT.

Algorithm 5 Intel MMX/SSE IDCT

```
pmaxsw  mm0, [CLIP_MIN] ;; compute first element
pminsw  mm0, [CLIP_MAX] ;; in order to free mm0
movq    [esi + 0], mm0  ; store x0 [0..3]
movq    mm0, [CLIP_MIN]  ; clip to -256
pmaxsw  mm1, mm0
pmaxsw  mm2, mm0
pmaxsw  mm3, mm0
pmaxsw  mm4, mm0
pmaxsw  mm5, mm0
pmaxsw  mm6, mm0
pmaxsw  mm7, mm0
movq    mm0, [CLIP_MAX]  ; clip to +255
pminsw  mm1, mm0
pminsw  mm2, mm0
pminsw  mm3, mm0
pminsw  mm4, mm0
pminsw  mm5, mm0
pminsw  mm6, mm0
pminsw  mm7, mm0
```

Although max and min can be synthesized through simpler operations (operations which are useful in their own right), the additional execution cost is simply too great to be practical. Rather than a single independent instruction, three dependent instructions are required. An arbitrary clamping operation can also be simulated with packed signed saturating addition. The representable range of a signed fixed point number (a bits to the left of the binary point, b bits to the right) is $-2^a \leq x \leq 2^a - 2^{-b}$. For example, if we need to limit a value, X, to the range $-j..+k$:

1. $(2^a - 2^b) - k \longrightarrow T_{pos}$
2. $X + T_{pos} \Longrightarrow X$
3. $X - T_{pos} \longrightarrow X$

These three steps limit X to $+K$. ($\Longrightarrow$ represents saturating overflow, where as a $\longrightarrow$ symbolizes modulo overflow.) Three more operations are required to limit X to the desired floor value:

1. $-2^a + j \longrightarrow T_{neg}$

2. $X + T_{neg} \Longrightarrow X$

3. $X - T_{neg} \longrightarrow X$

Architecturally, the implementation cost of max and min instructions should be low since the necessary comparators must already exist for saturating arithmetic. The only difference is that instead of comparing to a constant, a register value is used instead. An added advantage that we have found is that in many cases where comparisons are required, max and min instructions are sufficient.

Comparisons We have found that integer control flow instructions (e.g. packed comparisons) are seldom needed, except on architectures without max/min operations (e.g. Sun's VIS). We found one instance where a specialized floating point comparison was useful. In the project and clip test kernel, 3D objects are first mapped to 2D space through a matrix multiplication of 1x4 vectors and 4x4 matrices. Objects are then clipped to the viewable area to avoid unnecessary rendering. The code fragment listed in Algorithm 6 is computed for each vertex in a 3D scene every time a frame is rendered.

Algorithm 6 Clip Test and Project

```
FLOAT32 ex = vEye[i][0], ey = vEye[i][1], ez = vEye[i][2], ew =
vEye[i][3];
FLOAT32 cx = m0 * ex + m8 * ez, cy = m5 * ey + m9 * ez;
FLOAT32 cz = m10 * ez + m14 * ew, cw = -ez;
UINT8 mask = 0;
vClip[i][0] = cx; vClip[i][1] = cy; vClip[i][2] = cz; vClip[i][3] = cw;
if (cx > cw) mask |= CLIP_RIGHT_BIT;
else if (cx < -cw) mask |= CLIP_LEFT_BIT;
if (cy > cw) mask |= CLIP_TOP_BIT;
else if (cy < -cw) mask |= CLIP_BOTTOM_BIT;
if (cz > cw) mask |= CLIP_FAR_BIT;
else if (cz < -cw) mask |= CLIP_NEAR_BIT;
```

Motorola's AltiVec includes a specialized comparison instruction, `vcmpbfp`, which deals with boundary testing. This is done by testing all of the clip values in parallel to see if any clipping is needed, and branching to act as a fast out if no clipping is necessary. This technique is extremely effective because no clipping is the common case, with most vertices within the screen boundaries. An example of this is shown in the clip test kernel from Mesa's 3D rendering pipeline (Algorithm 6). For the Mesa "gears" application, the fast out case held true for 61946 of the 64560 (96.0%) clipping tests performed in our application run of 30 frames rendered at 1024x768 resolution.

Sum of Absolute Differences A sum of absolute differences (SAD) instruction operates on a pair of packed 8-bit unsigned input registers, summing the absolute differences between the respective packed elements in two registers and placing (or accumulating) the scalar sum in another register. The block match kernel is the only one in which sum of absolute differences (SAD) instructions is used. Algorithm 8 lists the core lines of the block match kernel utilized by MPEG-2 encoding. Block match sums the absolute differences between the corresponding pixels of two 16x16 macroblocks. The original application code also includes three other variations on

Algorithm 7 Motorola G4 Clip Test and Project

```
;;                          ; v1:  | cx | cy | cz | cw |
vspltw   v2, v1, 3          ; v2:  | cw | cw | cw | cw |
vcmpbfp. v3, v1, v2         ; v3:  bit mask of clip comparisons
;; set cr6 to 0x2 if all test values are within boundaries
li       r12, 0
mcrf     cr0, cr6
bc       COND_TRUE, 0x2, fast_out
vcmpgtsw v4, v0, v3         ; v4:  > test, - mask if clipping
vcmpgtsw v5, v3, v0         ; v5:  < test, + mask if clipping
vand     v4, v4, v27
vand     v5, v5, v28
vor      v4, v4, v5         ; v4:  | mask0| mask1| mask2| mask3|
vsldoi   v5, v4, v4, 8
vor      v4, v4, v5
vsldoi   v5, v4, v4, 4
vor      v4, v4, v5         ; v4:  | mask | mask | mask | mask |
vspltb   v4, v4, 15         ; v4:  | M| M|...| M| M| [0..15]
stvebx   v4, 0, r5          ; store mask to mask_temp
lbz      r12, 0(r5)         ; r12:  mask
lbz      r15, 0(r7)         ; r15:  clipMask[i]
or       r15, r15, r12
stb      r15, 0(r7)         ; r15:  clipMask[i] |= mask
or       r13, r13, r12      ; r13:  tmpOrMask |= mask
fast_out:
addi     r7, r7, 1          ; r7:  clipMask++ (pointer to UINT8)
and      r14, r14, r12      ; r14:  tmpAndMask &= mask
addic.   r3, r3, -1         ; n--
bc COND_FALSE, ZERO_RESULT, loop
```

block match which compute horizontal, vertical or both horizontal and vertical interpolation before calculating the sum of absolute differences.

Algorithm 8 Block Match

```
UINT8 blk_1[16][16]; UINT8 blk_2[16][16]; INT32 sad=0; INT32 diff;
for(j=0; j<h; j++)
 for(i=0; i<16; i++) {
  if ((diff = blk_1[j][i] - blk_2[j][i])<0)
    diff = -diff;
  sad+=diff;
 }
```

Although DEC's MVI extension is quite small (only 13 instructions), one of the few operations that DEC did include was SAD. DEC architects found (in agreement with our experience) that this operation provides the most performance benefit of all multimedia extension operations [Rubi96]. Intel's MMX, although a much richer set of instructions, did not include this operation (it *was* later included in both AMD's 3DNow!+ and Intel's SSE extensions to MMX). Sun's VIS also includes a sum of absolute differences instruction.

The Motorola G4 microprocessor was the only CPU in our survey which did not include some form of SAD operation, forcing us to synthesize the SAD operation from other instructions (Algorithm 9). Although Intel's SSE extension (see Algorithm 10) includes the `psadbw` instruction, this offers only a one cycle performance advantage when compared to the AltiVec implementation. In some ways this comparison is misleading since Intel's extension is 64-bits wide, while Motorola's is 128-bits; the question of performance should be absolute, not relative to Intel. To estimate the latency of a hypothetical SAD instruction for a 128-bit extension such as AltiVec, we examine the latency of this instruction on the other (64-bit) architectures:

Processor	Instruction	Latency:Throughput
Intel Pentium III	PSADBW	5 cycles : 1 every 2 cycles
AMD Athlon	PSADBW	3 cycles : 1 every 1 cycle
Sun UltraSPARC IIi	PDIST	3 cycles : 1 every 1 cycle
DEC Alpha 21264	PERR	2 cycles : 1 every 1 cycle

The latency of a 128-bit instruction would be higher than the 64-bit instructions listed in the table because this instruction requires a cascade of adders to sum (reduce) the differences between the elements. An N-bit SAD instruction ($M = N/8$) can be broken down into steps: 1) calculate M 8-bit differences, 2) calculate the absolute value of the differences, 3) perform $log_2 M$ cascaded summations. The architects of the 64-bit DEC MVI extension comment that a 3-cycle implementation of PERR would have been easily achievable, but in the end the architects achieved a more aggressive 2-cycle instruction [Carl97]. If a SAD operation were to be implemented in Motorola's AltiVec, we estimate it would have a latency of 4 cycles. This would certainly be a superior solution compared to the 9 cycle solution shown in Algorithm 9.

Algorithm 9 Motorola G4 Block Match - SAD portion (starting with vmaxub instruction) takes 9 cycles

```
;; v1:  block #1, line #1, pixels [0..F]
;; v4:  block #1, line #2, pixels [0..F]
;; v6:  block #2, line #1, pixels [0..F]
vavgub v1, v1, v4      ; v1:  vertically interpolated p0..pF
vmaxub v7, v1, v4      ; v7:  max [0..F]
vminub v8, v1, v4      ; v8:  min [0..F]
vsububs v7, v7, v8     ; v7:  abs_diffs [0..F]
vsum4ubs v31, v7, v31 ; v31:  | SAD_0 | SAD_1 | SAD_2 | SAD_3 |
vsumsws v31, v31, v0  ; v31:  |   0   |   0   |   0   |  SAD  |
```

Algorithm 10 Intel Pentium III Block Match - SAD portion (starting with first psadbw instruction) takes 8 cycles

```
;; mm1:  block #1, line #1, pixels [0..7]
;; mm5:  block #1, line #1, pixels [8..F]
;; mm3:  block #1, line #2, pixels [0..7]
;; mm4:  block #1, line #2, pixels [8..F]
;; mm2:  block #2, line #1, pixels [0..7]
;; mm6:  block #2, line #1, pixels [8..F]
pavgb mm1, mm3      ; mm1:  vertically interpolated p0..p7
pavgb mm5, mm4      ; mm5:  vertically interpolated p8..pF
psadbw mm1, mm2     ; mm1:  | 0 | 0 | 0 |SAD0| Pixels 0..7
psadbw mm5, mm6     ; mm5:  | 0 | 0 | 0 |SAD1| Pixels 8..F
paddd mm1, mm5      ; mm1:  | 0 | 0 | 0 | SAD|
```

Average. In addition to compute the sum of absolute differences, half-pixel interpolation, for which MPEG-2 encoding offers three varieties, is also important; vertical interpolation is shown in Algorithms 9 and 10. Interpolation is done by averaging a set of pixel values with pixels offset by one horizontally, vertically or both. The original C MPEG-2 code first performs the interpolation, and then computes the sum of absolute differences on the result. SIMD interpolation can be performed through 8-bit unsigned average instructions (again see Algorithms 9 and 10). DEC's MVI extension does not include a packed average instruction, but a similar interpolation operation can be done by averaging the result of several SAD operations using scalar integer arithmetic (since the result of a SAD instruction is a scalar value).

Integer average operations were only used in the block match kernel. This kernel operates on 8-bit unsigned values, so this is the only type of "average" instruction that was useful within our workload.

High Latency Function Approximation. Applications such as 3D rendering have kernels which use floating point mathematical functions, such as reciprocal and square-root, that are very high latency. On some architectures these scalar functions are computed in software, while others have hardware instructions. Full IEEE compliant operations return 24-bits of mantissa. The computation of these functions is iterative, so the number of bits of precision returned is directly proportional to an operation's latency. It is for this reason that all of the SIMD floating point extensions (AMD's 3DNow!, Intel's SSE and Motorola's AltiVec) include approximation instructions for $\frac{1}{x}$ and $\frac{1}{\sqrt{x}}$. These are typically implemented as hardware lookup tables, returning k-bits of precision. In Intel's SSE, for example, approximate reciprocal (`rcp`) and reciprocal square root (`rsqrt`) return 12-bits of mantissa. Motorola's AltiVec also returns 12-bits of precision for both the reciprocal and reciprocal square root approximation instructions.

In the transform and normalize kernel, graphics primitives are transformed to the viewer's frame of reference through matrix multiplications. The code shown in Algorithm 11 is computed for each vertex in a 3D scene.

Algorithm 11 Transform and Normalize

```
FLOAT64 tx, ty, tz, len, scale;
FLOAT32 ux = u[i][0],  uy = u[i][1],  uz = u[i][2];
tx = ux * m[0] + uy * m[1] + uz * m[2];
ty = ux * m[4] + uy * m[5] + uz * m[6];
tz = ux * m[8] + uy * m[9] + uz * m[10];
len = sqrt( tx*tx + ty*ty + tz*tz );
scale = (len>1E-30) ?  (1.0 / len) :  1.0;
v[i][0] = tx * scale; v[i][1] = ty * scale;
v[i][2] = tz * scale;
```

The transform kernel has at its heart a floating point reciprocal square root operation ($\frac{1}{\sqrt{x}}$). One unique aspect of Intel's SSE instruction set is that not only does it include 22-bit precise (mantissa) approximations of $\frac{1}{x}$ and $\frac{1}{\sqrt{x}}$, but it also includes full precision (24-bits of mantissa) versions of division and $\sqrt{x}$. Of course, this added precision comes at a price - namely much higher latency (their full precision instructions are not pipelined) than the pipelined 22-bit approximations derived from processor internal lookup tables. All of the floating point multimedia extension vendors, including Intel, point out the Newton-Raphson method for improving the accuracy of approximations through specially derived functions. In the case of $\frac{1}{\sqrt{x}}$ it is possible to iteratively increase the precision of an initial approximation through the equation:

$$x_1 = x_0 - (0.5 \cdot a \cdot x_0^3 - 0.5 \cdot x_0) = 0.5 \cdot x_0 \cdot (3.0 - a \cdot x_0^2) \qquad (1)$$

Employing an approximation instruction in conjunction with the Newton-Raphson method to achieve full precision is actually faster than the full precision version of the instruction that Intel provides. Compare the code fragments in Algorithms 12 and 13, which have execution times of 25 vs.

36 cycles. One iteration of the Newton-Raphson method is enough to improve a 22-bit approximation to the full 24-bit precision of IEEE single precision.

Algorithm 12 Intel Approximated Square Root - 25 cycles

```
; xmm3:  tx^2+ty^2+tz^2 [0..3] (len=sqrt(tx^2+ty^2+tz^2))
; xmm7:  0.5 [0..3]
; xmm5:  3.0 [0..3]
rsqrtps  xmm4, xmm3        ; xmm4:  rsqrtps(a)
movaps   xmm6, xmm3
mulps    xmm6, xmm4        ; xmm6:  a*rsqrtps(a)
mulps    xmm6, xmm4        ; xmm6:  a*rsqrtps(a)*rsqrtps(a)
mulps    xmm4, xmm7        ; xmm4:  0.5*rsqrtps(a)
subps    xmm5, xmm6        ; xmm5:  3.0 - a*rsqrtps(a)*rsqrtps(a)
mulps    xmm4, xmm5        ; xmm4:  |1/len1|1/len2|1/len1|1/len0|
;; sqrt(a) = a*(1/sqrt(a))
mulps    xmm3, xmm4        ; xmm3:  | len3 | len2 | len1 | len0 |
```

Algorithm 13 Intel Full Precision Square Root - 36 cycles

```
; xmm3:  tx + ty + tz [0..3] (len=sqrt(tx+ty+tz))
; xmm3:  a [0..3]
sqrtps   xmm3, xmm3        ; xmm3:  sqrt(a)
```

The added cost of the Newton-Raphson method is of the additional register space needed to hold intermediate values and constants. AMD's 3DNow! extension circumvents this by including instructions to internally perform the Newton-Raphson method, rather than having the programmer implement it (Algorithm 14). The only odd thing about AMD's reciprocal square root instructions are that they are actually scalar; they only produce one result value, based on the lower packed element.

Algorithm 14 AMD Approximated Square Root - 20 cycles

```
; mm3:  tx^2+ty^2+tz^2 [0] (len=sqrt(tx^2+ty^2+tz^2))
pfrsqrt          mm4, mm7      ; mm4: |~1/len0 |~1/len0 |
movq             mm5, mm4
pfmul            mm4, mm4
pfrsqit1         mm4, mm7
pfrcpit2         mm4, mm5      ; mm4: | 1/len0 | 1/len0 |
pfmul            mm7, mm4      ; mm7: |  len0  |  len0  |
```

5.3.2 Exceptions

Techniques for handling exceptions that occur during SIMD processing are very similar to those employed when dealing with packed overflow. Checking result flags or generating an exception from a packed operation requires considerable time to determine which packed element caused the problem. In most cases where an exception might be raised it is possible to fill in a value which will give reasonable results for most applications. This speeds execution because no error condition checking need be done, and is is similar to saturating integer arithmetic where maximum or minimum result values are substituted rather than checking for and reporting positive or negative overflow. Both AMD's 3DNow! and Motorola's AltiVec extensions do not implement IEEE compliant floating point exceptions. Only Intel's SSE implements full IEEE compliant SIMD floating point exceptions, and includes a control/status register (`MXCSR`) to mask or unmask packed floating point numerical exceptions.

5.3.3 Floating Point Rounding

Intel's SSE offers two modes of rounding: IEEE compliant and another, faster, flush to zero (FTZ) mode. Flush to zero (FTZ) clamps to a minimum representable result in the event of underflow (a number too small to be represented in single precision floating point). Fully compliant IEEE floating point supports four rounding modes. Most real time 3D applications use the FTZ rounding mode since they are not particularly sensitive to a slight loss in precision [Thak99]. 3DNow! supports only truncated rounding (round to zero). All of Motorola's AltiVec floating point arithmetic instructions use the IEEE default rounding mode of round to nearest. The IEEE directed rounding modes are not provided.

5.3.4 Type Conversion

Width promotion is the expansion of an N-bit value to some larger width. For unsigned fixed point numbers this requires zero extension or filling any additional bits with zeros. Zero extension is usually not specified as such in a multimedia architecture because it overlaps in functionality with data rearrangement instructions such as unpack or merge. If packed values are merged with another register which has been zeroed prior to merging the result is zero extension. Signed element unpacking is not as simple, but is rarely supported directly by hardware; only the AltiVec instruction set includes it. It can be synthesized with multiplication by one since a multiplication yields a result that is the overall width of both its operands.

Video and imaging algorithms use an 8-bit unsigned data type. Audio and speech algorithms, on the other hand, typically employ signed 16-bit values, but because multiplication by a fractional fixed point constant is a common operation, these values are often unpacked as a natural consequence of computation. So, although a signed unpack operation would likely be faster than multiplication by 1, it is seldom necessary to resort to this in practice.

All data types that occur in multimedia should be supported for packing and unpacking even for those widths not directly supported by arithmetic operations. It should always be possible to convert to a width that is supported for computation. Although we do not otherwise examine HP's MAX-1/MAX-2 extensions, as no hardware employing them was available to us at the time of this work, they are good examples of where not following this guidline can cause problems. We have noted the importance of the 16-bit data width. HP's MAX-1/MAX-2 instruction sets only support operations on 16-bit wide values. Partitioned 8-bit operations were considered, but rejected due to insufficient precision. Wider packed data types (e.g. 32-bit) were not included due to insufficient parallelism. What this approach overlooks is that fact that even though many intermediate computations require greater precision than 8-bits, many types of video and imaging data

are stored this way in existing multimedia file formats. Thus, packing and unpacking to and from 8-bit precision is a very common operation, which is not supported in hardware, making HP's extensions inefficient at processing this type of data.

5.3.5 Data Rearrangement

SIMD instructions perform the same operation on multiple data elements. Because all of the data within a register must be treated identically, the ability to efficiently rearrange data bytes within and between registers is critical for performance. We will refer to these types of operations as "data communication" instructions. *Interleave* instructions (also referred to as *mixing*, *unpacking* or *merging*) merge alternate data elements from the upper or lower half of the elements in each of two source registers. *Align* or *rotate* operations allow for arbitrary byte-boundary data realignment of the data in two source registers; essentially a shift operation that is done in multiples of 8-bits at a time. Both interleave and align type operations have hard coded data communication patterns. *Insert* and *extract* operations allow for a specific packed element to be extracted as a scalar or a scalar value to be inserted to a specified location. *Shuffle* (also called *permute*) operations allow greater flexibility than those operations with fixed communication patterns, but this added flexibility requires that the communication pattern be specified either in a third source register or as an immediate value in part of the instruction encoding.

The sufficiency of simpler data communication operations is to some degree dependent on the vector length employed. For example, 128-bit AltiVec vectors contain up to sixteen elements, while a shorter extension such as Intel's 64-bit MMX contain at most eight of the same type of element. This means that simple data rearrangement operations (e.g. merge) cover a relatively larger fraction of all possible mappings in the case of the shorter vector length. [Lee00] presents a novel set of simple data communication primitives which can perform all 24 permutations of a 2x2 matrix in a single cycle on a processor with dual data communication functional units. This is useful because any larger data communication problem can be decomposed into 2x2 matrices. Although this approach might make some very complex data communication patterns slow to compute, we have found that most multimedia algorithms have patterns which are relatively simple. Because of this we endorse [Lee00]'s technique for covering the data communication needs of multimedia applications.

A related class of instructions that the AltiVec extension included, that was quite useful, was a set of "splat" instructions, which place either an immediate scalar or specified element from a source register into every element of the destination register. This was very useful when constants were required; on other architectures it is necessary to statically store these types of values in memory, and then load them to a register when required.

5.3.6 Prefetching

Prefetching is a hardware or software technique which tries to predict data access needs in advance, overlapping memory access with useful computation. Although we will not otherwise examine the performance implications of prefetch instructions (which would be a useful extension to this study), we mention them briefly because they are often a part of multimedia instruction sets due to the highly predictable nature of data accesses in multimedia applications.

Software prefetch instructions are used to fetch data into the cache from main memory without blocking true load/store instruction accesses. Determining the ideal location for prefetch instructions in a piece of code depends on many architectural parameters. Unfortunately, these include such things as the number of CPU clocks for memory latency and the number of CPU clocks to transfer a cache line, which are both highly machine dependent and not readily available to the programmer.

Rather than issuing an explicit prefetch instruction for each desired data prefetch, Motorola's AltiVec uses a single *data stream touch instruction* (`dst`) which indicates the memory sequence or pattern that is likely to be accessed. We will refer to this hybrid of hardware and software prefetching as *software directed prefetching* to indicate that a separate prefetch instruction need not be issued for each data element. A data stream is defined by a sequence starting address, size of each unit (up to 32 128-bit blocks), total number of units (up to 256), bytes between units (-32768..+32767) and a 2-bit ID tag for the stream. Hardware optimizes the number of cache blocks to prefetch so it is not necessary for the programmer to know the parameters of the cache system. A stream is fetched either until all of the requested blocks have been brought into the cache or another `dst` instruction is issued with the same tag ID. The stream construct eliminates the instruction issue overhead as well as the problem of determining the optimal prefetch distance.

5.4 Bottlenecks and Unnecessary Features

In this section we discuss those features which appear in multimedia instruction sets, do not appear to be useful, and are not "free"; i.e. they aren't a low (or no) cost side effect of some other useful feature.

Instruction Primitives The VIS instruction set does not include full 16-bit multiply instructions. It instead offers multiplication primitives, the results of which must be combined through addition (see Algorithms 15 and 16).

Algorithm 15 Sun VIS 16-bit x 16-bit →16-bit Multiply

```
fmul8sux16   %f0, %f2, %f4
fmul8ulx16   %f0, %f2, %f6
fpadd16      %f4, %f6, %f8
```

The Mix stereo kernel is a good example of the high cost of synthesizing needed instruction functionality from other primitives. Audio mixing consists of multiplying a vector of

Algorithm 16 Sun VIS 16-bit x 16-bit →32-bit Multiply

```
fmuld8sux16  %f0, %f2, %f4
fmuld8ulx16  %f0, %f2, %f6
fpadd32      %f4, %f6, %f8
```

count input signals (*sp*[]) by a vector of mixing coefficients $chan_1, chan_2$ and summing the result (Algorithm 17). In the Timidity MIDI music synthesis application, fixed point integer computations are used to mix the various signed 16-bit instrument sounds into a 32-bit output buffer.

Algorithm 17 Mix Stereo

```
INT16 **sp_p; INT32 **lp_p; INT32 count; INT32 chan_1; INT32 chan_2;
  INT16 s, *sp = *(sp_p); INT32 *lp = *(lp_p);
  while (count--) {
    s = *sp++;
    *lp++ += s*chan_1;
    *lp++ += s*chan_2;
  }
  *(sp_p) = sp;
  *(lp_p) = lp;
}
```

Comparing the code snippet in Algorithm 18 to Algorithm 19 we can see that Sun's approach of synthesizing functionality from primitives (especially in the case of synthesizing a 16-bit merge) is much more costly than using a single instruction.

Algorithm 18 Intel Mix

```
movq    mm0, [esi]           ; mm0: |  s3 |  s2 |  s1 |  s0 |
movq    mm1, [edi]           ; mm1: |   lp1    |   lp0    |
movq    mm2, [edi + 8]       ; mm2: |   lp3    |   lp2    |
pshufw  mm5, mm0, 00000000b  ; mm5: |  s0 |  s0 |  s0 |  s0 |
pshufw  mm6, mm0, 01010101b  ; mm6: |  s1 |  s1 |  s1 |  s1 |
pmaddwd mm5, mm7             ; mm5: | s0*right  |  s0*left  |
pmaddwd mm6, mm7             ; mm6: | s1*right  |  s1*left  |
paddd   mm1, mm5             ; mm1: |   lp1'    |   lp0'    |
paddd   mm2, mm6             ; mm2: |   lp3'    |   lp2'    |
```

The reason that the architects of VIS divided up 16-bit multiplication in this way was to decrease die area. Not providing a full 16x16 multiplier subunit cut the size of the arrays in half [Trem96b]. Unfortunately, dividing an operation into several instructions (which are not otherwise useful in and of themselves) increases register pressure, decreases instruction decoding bandwidth and creates additional data dependencies. Splitting SIMD instructions (which have been introduced for their ability to extract data parallelism) can actually cripple a superscalar processor's ability to extract instruction level parallelism. A multi-cycle operation can be a better solution than a multi-instruction operation because instruction latencies can be transparently upgraded in future processors, while poor instruction semantics can not be repaired without adding new instructions.

Unused High Latency Approximation Instructions Floating point approximation of $\frac{1}{x}$ instructions, although available on several platforms, did not find application in any of the kernels we studied. AltiVec also includes approximate log_2 and exp_2 instructions, which find application in the lighting stage of a 3D rendering pipeline; this is currently handled by 3D accelerator cards, and not the CPU.

Algorithm 19 Sun Mix

```
wr        %g0, 6, %gsr     !set alignment for right shift by 16
ld        [%l1 +  0], %f16 !%f16:|    lp0    |XXXXXXXXXXX|
ld        [%l1 +  4], %f17 !%f17:|    lp0    |    lp1    |
ld        [%l1 +  8], %f18 !%f18:|    lp2    |XXXXXXXXXXX|
ld        [%l1 + 12], %f19 !%f19:|    lp2    |    lp3    |
!  simulate 16-bit merge
fpmerge   %f28, %f28, %f4  !%f4: |s0_|s0_|_s0|_s0|s1_|s1_|_s1|_s1|
fpmerge   %f29, %f29, %f10 !%f10:|s2_|s2_|_s2|_s2|s3_|s3_|_s3|_s3|
faligndata %f4, %f4, %f6   !%f6: |_s1|_s1|s0_|s0_|_s0|_s0|s1_|s1_|
faligndata %f10, %f10, %f12 !%f12:|_s3|_s3|s2_|s2_|_s2|_s2|s3_|s3_|
fpmerge   %f6, %f4, %f8    !%f8: |XXX|XXX|XXX|XXX|   s0  |   s0  |
fpmerge   %f12, %f10, %f14 !%f14:|XXX|XXX|XXX|XXX|   s2  |   s2  |
fpmerge   %f7, %f5, %f2    !%f2: |XXX|XXX|XXX|XXX|   s1  |   s1  |
fpmerge   %f13, %f11, %f26 !%f26:|XXX|XXX|XXX|XXX|   s3  |   s3  |
fsrc1s    %f9, %f2         !%f2: |  s0 |  s0 |  s1 |  s1 |
fsrc1s    %f15, %f26       !%f26:|  s2 |  s2 |  s3 |  s3 |
!  simulate 16x16 -> 32-bit multiply
fmuld8sux16 %f0, %f2, %f4
fmuld8ulx16 %f0, %f2, %f6
!  simulate 16x16 -> 32-bit multiply
fmuld8sux16 %f0, %f3, %f8
fmuld8ulx16 %f0, %f3, %f10
fpadd32   %f4, %f6, %f4    !  %f4:  |  s0*chan1 |  s0*chan2 |
fpadd32   %f8, %f10, %f6   !  %f6:  |  s1*chan1 |  s1*chan2 |
fpadd32   %f16, %f4, %f16  !  %f16: |    lp0'   |    lp1'   |
fpadd32   %f18, %f6, %f18  !  %f18: |    lp2'   |    lp3'   |
```

Unused Pixel Conversion Instructions Motorola's AltiVec extension includes pixel pack (`vpkpx`) and pixel unpack (`vupkhpx`, `vupklpx`) instructions for converting between 32-bit true color and 16-bit color representations. These did not find application within the BMKL, although it is possible that they might be of utility in situations where AltiVec needs to operate on 16-bit color data; many video games use 16-bit textures, for example.

Unused Memory Access Instructions Sun's VIS includes two sets of instructions for accelerating multimedia operations with sophisticated memory addressing needs. The first, `edge8`, `edge16`, and `edge32`, produce bit vectors to be used in conjunction with partial store instructions to deal with the boundaries in 2D images. The second group of addressing instructions include `array8`, `array16` and `array32` which find use in *volumetric imaging* (the process of displaying a two dimensional slice of a three dimensional data). An array instruction converts (x, y, z) coordinates into a memory address. The Berkeley multimedia workload does not include any volumetric imaging applications, so it is unsurprising that these instructions found no utility in our workload.

Singular, Highly Utilized Resources Although we usually think of SIMD architectures as extracting data level parallelism, all of the implementations of the instruction sets we have examined are also superscalar, with multiple parallel SIMD functional units. In fact, unless the SIMD vector length is long enough to hold the entire data set being operated on, there is almost always the potential to extract instruction level parallelism as well. In coding the kernels with Sun's VIS extension, it became clear that instruction level parallelism was being compromised by the over utilized graphics status register (GSR).

Sun's VIS architecture does not include partitioned shift instructions, the GSR has a 3-bit `addr_offset` field which is used implicitly for byte granularity alignment, as well

as a 4-bit `scale_factor` field for packing/truncation operations. The VIS GSR is a serializing bottleneck because any time packing or alignment functionality is needed, it must be pushed through the GSR. Because VIS lacks partitioned shift operations, we found ourselves synthesizing such operations with the packing and alignment operations where no other algorithmic path was possible. Even with careful planning of packing and alignment operations it was often necessary to write to the GSR several times in each iteration of the loops of our multimedia kernels. The serializing effect of this singular resource prevented VIS operations from proceeding at the fullest possible degree of parallelism.

5.5 Alignment and Memory Traffic

Factors such as register file geometry (the number of registers and their width), data path location (pre-existing integer or floating point, or separate) and alignment issues are reflected in the uncached memory traffic - the data accesses as seen by the L1 data cache. Table 3 lists the average number of bytes loaded or stored per function call. This was computed by multiplying the number of dynamic load and store instructions executed by their widths in bytes.

From Table 3 we can see that Motorola's and Sun's implementations sometimes seem to transfer (load and store) more bytes in each function call than the AMD, DEC or Intel implementations of the same kernel. We would expect the Motorola and Sun implementations to spill registers to memory less frequently due to their larger register files (on average we see that the register file geometry does affect memory traffic as we might expect). This surprising result is actually an artifact of how we computed memory traffic, rather than an indication of an architectural shortcoming. Dynamic instruction counts alone are not a completely accurate predictor of actual memory traffic, since some of the architectures (AMD, DEC, Intel) support unaligned loads (hiding some loads issued by the hardware which handles unaligned memory access in the CPU) and the rest do not. Hardware support to efficiently handle memory accesses that are not aligned are expensive in both area and timing [Thak99].

Transparently Forced Alignment The AltiVec instruction set architecture does not provide for alignment exceptions when loading and storing data. Alignment is maintained by forcing the lower four bits of any address to be zero. This is transparent to the programmer, so the programmer is responsible for guaranteeing alignment, otherwise incorrect data may be loaded or stored. We believe it is better that performance and correctness issues due to alignment be made explicit. The loading of incorrect data due to a mistaken assumption about alignment would be an extremely difficult bug to track down.

5.6 Overall Instruction Mix

Table 4 shows what types of instructions comprised the total mix of dynamic instructions executed by each architecture. Counts are for the code within the kernels only, and do not include instructions from the rest of each application. Control state instructions are those which interact with special purpose machine registers (e.g. the GSR on Sun's VIS). Branches and other data dependent codes such as conditional moves are categorized as "control flow" type instructions.

We can see that SIMD kernels utilize a significant number of scalar operations for pointers, loop variables and clean up code; evidence that sharing the integer data path is not a good idea. Intel's SSE is a 128-bit wide extension, as compared to AMD's 64-bit wide 3DNow!, explaining why Intel's overall instruction count is lower by about 1 billion instructions. The same reasoning applies to Motorola's G4 which has the 128-bit wide (both packed integer and floating point) AltiVec extension. DEC's bloated instruction count is due to the fact that their MVI extension is very limited in functionality (13 instructions in total), and so many operations need to be done with scalar instructions.

Data communication operations represent the overhead necessary to put data in a format amenable to SIMD operations. Ideally, these types of instructions should make up as small a part of the dynamic instruction mix as possible. Table 4 reveals that the Motorola AltiVec extension executes the largest percentage of these instructions. This is due to two factors: 1) the wider register width means that it is less likely that the data is arranged correctly as first loaded from memory and 2) data communication operations are used by AltiVec to simulate unaligned access to memory.

6 Performance Comparison

In order to establish a base line for performance, the average execution time of the C and SIMD assembly versions of the BMKL were measured on each platform (Table 5). A speedup from 2x-5x was typical, although some kernels were sped up considerably more than this. Note that the C compiler for the Apple system running a pre-release version of OS X (developer's preview 3) is known to be weak, making the speedup of AltiVec over C look unrealistically good.

All of our performance measurements utilize the metric of time rather than cycle or instruction counts. If all of the architectures in our study had equal cycle times, then comparing cycle times would be valid, since time, in that case, would simply be proportional. This is not the case, since the Sun UltraSPARC IIi used in our study has a 360 MHz clock, while the remainder of the chips are 500 MHz parts. Instruction counts are not valid measure for cross architectural comparisons, as each instruction set does not necessarily do the same amount of work (computation) in the same number of instructions, nor take a the same amount of time to perform similar instructions.

To measure how well or how poorly a given platform performs relative to the competition we use the metric of *percent deviation from the mean*:

$$\%DM = 100 \cdot \left(\frac{\bar{t} - t_i}{\bar{t}}\right) \quad (2)$$

where t_i is the time taken on platform i, and $\bar{t}$ is the average performance (time) across all of the platforms examined for

	AMD		DEC		Intel		Motorola		Sun		Average
Register File Geometery	8x64b (FP)		31x64b (Int)		8x64b (FP), 8x128b		32x128b		32x64b		
Add Block	296	*2.8%*	249	*18.2%*	290	*4.7%*	431	*-41.7%*	256	*16.0%*	304
Block Match	356	*27.7%*	572	*-16.2%*	356	*27.8%*	730	*-48.2%*	448	*9.0%*	492
Clip Test & Project	5,261	*-10.8%*	*6,755*	*-42.3%*	2,840	*40.2%*	2,139	*54.9%*	*6,735*	*-41.9%*	4,746
Color Space	14,517,360	*-101.9%*	2,073,744	*71.2%*	14,517,360	*-101.9%*	2,419,504	*66.3%*	2,419,296	*66.3%*	7,189,453
DCT	2,424	*52.4%*	*19,008*	*-273.1%*	2,420	*52.5%*	304	*94.0%*	1,320	*74.1%*	5,095
FFT	99,797	*12.0%*	*171,163*	*-50.9%*	95,117	*16.1%*	118,560	*-4.5%*	*82,408*	*27.3%*	113,409
IDCT	1,640	*-60.6%*	*682*	*33.2%*	1,640	*-60.6%*	320	*68.7%*	824	*19.3%*	1,021
Max Value	852	*20.6%*	782	*27.1%*	852	*20.6%*	2,098	*-95.5%*	783	*27.1%*	1,073
Mix Stereo	823	*-17.5%*	*503*	*28.2%*	823	*-17.5%*	505	*27.9%*	849	*-21.2%*	701
Quantize	4,733	*45.6%*	*7,391*	*15.1%*	4,675	*46.3%*	21,098	*-142.5%*	*5,609*	*35.5%*	8,701
Short Term Anal. Filter	8,584	*-105.0%*	*3,056*	*27.0%*	8,584	*-105.0%*	272	*93.5%*	440	*89.5%*	4,187
Short Term Synth. Filter	7,100	*-93.3%*	*3,448*	*6.1%*	7,100	*-93.3%*	274	*92.5%*	441	*88.0%*	3,673
Subsample Horizontal	13,685,804	*-67.3%*	*2,945,448*	*64.0%*	13,685,892	*-67.3%*	5,642,508	*31.0%*	4,930,600	*39.7%*	8,178,050
Subsample Vertical	8,300,932	*-88.5%*	1,123,328	*74.5%*	8,301,052	*-88.5%*	2,137,092	*51.5%*	2,160,040	*51.0%*	4,404,489
Synthesis Filter	4,080	*-2.2%*	*4,136*	*-3.6%*	4,144	*-3.8%*	3,273	*18.0%*	*4,328*	*-8.4%*	3,992
Xform & Normalize	3,037	*-34.6%*	*1,976*	*12.5%*	2,162	*4.2%*	1,659	*26.5%*	*2,451*	*-8.6%*	2,257
Average	**2,290,192**	***-26.3%***	**397,640**	***-0.6%***	**2,289,707**	***-20.4%***	**646,923**	***18.3%***	**601,052**	***28.9%***	

Table 3: **SIMD Kernel Memory Traffic** - data bytes transferred per call as seen by the L1 cache (in other words, uncached memory traffic) are listed in the left subcolumns, with the percent deviation from the mean values (%DM) given in the right subcolumns. Values in *italics* indicate kernels which are implemented in C due to lacking SIMD instruction set functionality.

	AMD Athlon		DEC 21264A		Intel Pentium III		Motorola G4		Sun UltraSPARC IIi	
Int Load/Store	2.41E+08	*(5.4%)*	1.51E+09	*(20.3%)*	2.19E+08	*(6.2%)*	1.09E+08	*(2.4%)*	1.47E+08	*(2.6%)*
Int Arithmetic	5.43E+08	*(12.1%)*	2.18E+09	*(29.2%)*	4.34E+08	*(12.3%)*	8.65E+08	*(19.1%)*	1.49E+09	*(26.6%)*
Int Control Flow	5.75E+08	*(12.8%)*	8.41E+08	*(11.3%)*	4.73E+08	*(13.4%)*	6.26E+08	*(13.8%)*	5.20E+08	*(9.3%)*
Int Data Communication	1.03E+08	*(2.3%)*	4.98E+08	*(6.7%)*	7.50E+07	*(2.1%)*	4.75E+07	*(1.1%)*	0.00E+00	*(0.0%)*
FP Load/Store	1.42E+08	*(3.2%)*	8.36E+08	*(11.2%)*	1.40E+08	*(4.0%)*	3.35E+08	*(7.4%)*	4.48E+08	*(8.0%)*
FP Arithmetic	0.00E+00	*(0.0%)*	1.34E+09	*(18.1%)*	0.00E+00	*(0.0%)*	1.04E+08	*(2.3%)*	4.69E+08	*(8.4%)*
FP Control Flow	0.00E+00	*(0.0%)*	1.05E+07	*(0.1%)*	0.00E+00	*(0.0%)*	0.00E+00	*(0.0%)*	3.50E+08	*(6.2%)*
FP Data Communication	0.00E+00	*(0.0%)*	0.00E+00	*(0.0%)*	0.00E+00	*(0.0%)*	8.73E+05	*(0.0%)*	0.00E+00	*(0.0%)*
SIMD Load/Store	8.70E+08	*(19.4%)*	0.00E+00	*(0.0%)*	7.40E+08	*(21.0%)*	8.12E+08	*(17.9%)*	6.56E+08	*(11.7%)*
SIMD Data Communication	5.50E+08	*(12.3%)*	0.00E+00	*(0.0%)*	4.43E+08	*(12.6%)*	7.67E+08	*(17.0%)*	5.44E+08	*(9.7%)*
SIMD Int Arithmetic	5.61E+08	*(12.5%)*	2.28E+08	*(3.1%)*	6.62E+08	*(18.8%)*	5.36E+08	*(11.8%)*	7.61E+08	*(13.6%)*
SIMD Int Control Flow	0.00E+00	*(0.0%)*	0.00E+00	*(0.0%)*	0.00E+00	*(0.0%)*	5.23E+03	*(0.0%)*	1.97E+08	*(3.5%)*
SIMD FP Arithmetic	8.95E+08	*(19.9%)*	0.00E+00	*(0.0%)*	3.27E+08	*(9.3%)*	3.12E+08	*(6.9%)*	0.00E+00	*(0.0%)*
SIMD FP Control Flow	2.91E+05	*(0.0%)*	0.00E+00	*(0.0%)*	1.46E+05	*(0.0%)*	8.12E+04	*(0.0%)*	0.00E+00	*(0.0%)*
Control State	8.24E+06	*(0.2%)*	0.00E+00	*(0.0%)*	8.35E+06	*(0.2%)*	1.25E+07	*(0.3%)*	2.16E+07	*(0.4%)*
Total	**4.49E+09**	***100%***	**7.45E+09**	***100%***	**3.52E+09**	***100%***	**4.53E+09**	***100%***	**5.60E+09**	***100%***

Table 4: **Overall Instruction Mix** - counts are listed with percentages in parentheses. Control state instructions are those which interact with special purpose registers. Control flow instructions include branches as was as conditional moves and comparisons.

a particular kernel. This metric indicates how much better or worse than average, and provides a normalized result for computing the average improvement or degradation in performance. Table 6 lists the average %DM (the arithmetic average of the %DM values for all of the kernels on a given platform).

The algorithm employed by the original MPEG-2 encoder DCT code (Algorithm 3) is not very efficient - it uses double precision floating point where fixed point integer should provide sufficient precision (and is typically faster on most architectures) [IEEE91]. Because the DCT and IDCT algorithms are of the same computational order of magnitude it might seem strange that they demonstrate such different performance improvements (Table 5). Unlike the original forward DCT code which was computed in floating point, the scalar inverse DCT source code (Algorithm 4) is written in fixed point integer. Thus it is much more directly comparable to our fixed point integer SIMD implementations.

The fact that the original C DCT algorithm is floating point and the SIMD implementations are fixed-point integer meant that it was not appropriate to use the original code as the "solution" for the DEC Alpha architecture, which did not provide

	AMD Athlon			DEC Alpha 21264			Intel Pentium III			Motorola G4			Sun UltraSPARC III			Arithmetic Mean		
Add Block	(9.6x)	1,087	*114*	(1.9x)	669	*350*	(6.4x)	969	*152*	(2.5x)	1,054	*423*	(4.9x)	1,662	*342*	(5.0x)	1,088	276
Block Match	(7.0x)	1,801	*257*	(2.6x)	979	*379*	(4.9x)	1,855	*381*	(3.4x)	1,466	*437*	(2.8x)	2,408	*852*	(4.1x)	1,702	461
Clip Test & Project	(1.6x)	7,374	*4,559*	(1.0x)	8,591	*8,591*	(1.3x)	7,938	*6,336*	(2.6x)	8,906	*3,427*	(1.0x)	12,222	*12,222*	(1.5x)	9,006	7,027
Color Space	(9.4x)	122,103,855	*12,924,258*	(1.7x)	22,399,926	*13,043,849*	(5.6x)	60,141,946	*10,803,618*	(4.5x)	58,043,968	*12,959,530*	(2.7x)	62,015,071	*22,994,253*	(4.8x)	64,940,953	14,545,102
DCT**	(16.9x)	24,332	*1,438*	(19.7x)	12,475	*632*	(31.1x)	38,192	*1,228*	(36.6x)	33,176	*907*	(9.2x)	30,922	*3,353*	(22.7x)	27,819	1,512
FFT	(2.0x)	125,484	*63,446*	(1.0x)	67,321	*67,321*	(1.7x)	147,047	*84,661*	(2.0x)	232,272	*116,828*	(1.0x)	111,258	*111,258*	(1.5x)	136,677	88,703
IDCT	(3.6x)	2,999	*827*	(1.0x)	1,276	*1,276*	(2.8x)	2,772	*993*	(3.9x)	3,326	*847*	(1.5x)	3,782	*2,508*	(2.6x)	2,831	1,290
Max Value	(4.0x)	2,407	*604*	(1.0x)	1,143	*1,168*	(3.8x)	2,536	*669*	(6.6x)	4,110	*618*	(0.7x)	6,165	*8,715*	(3.2x)	3,272	2,355
Mix Stereo	(2.8x)	956	*341*	(1.0x)	616	*616*	(2.0x)	1,065	*528*	(3.8x)	1,710	*446*	(2.1x)	3,550	*1,716*	(2.3x)	1,579	729
Quantize	(1.8x)	34,233	*18,557*	(1.0x)	19,549	*19,549*	(2.5x)	37,100	*15,010*	(2.1x)	29,488	*14,335*	(1.0x)	34,928	*34,928*	(1.7x)	31,059	20,476
Short Term Analysis Filter	(1.5x)	15,268	*9,887*	(1.0x)	11,120	*11,120*	(1.5x)	16,129	*11,003*	(6.5x)	24,156	*3,729*	(2.5x)	36,773	*15,009*	(2.6x)	20,689	10,150
Short Term Synthesis Filter	(2.9x)	21,849	*7,425*	(1.0x)	19,208	*19,208*	(7.6x)	43,582	*5,759*	(6.5x)	23,721	*3,672*	(3.2x)	48,660	*15,202*	(4.2x)	31,404	10,253
Subsample Horizontal	(1.4x)	15,977,835	*11,337,253*	(1.0x)	9,210,928	*9,264,521*	(1.3x)	15,777,956	*11,714,792*	(1.6x)	17,472,826	*10,849,858*	(0.9x)	17,707,829	*19,521,092*	(1.3x)	15,229,475	12,537,503
Subsample Vertical	(2.4x)	25,517,638	*10,791,669*	(3.3x)	14,680,887	*4,487,870*	(2.2x)	21,969,585	*10,049,618*	(9.6x)	29,349,708	*3,070,046*	(3.8x)	39,872,378	*10,457,565*	(4.2x)	26,278,039	7,771,354
Synthesis Filter	(2.8x)	7,308	*2,585*	(1.0x)	4,900	*4,900*	(1.8x)	8,349	*4,718*	(1.8x)	4,917	*2,727*	(1.0x)	7,138	*7,138*	(1.7x)	6,522	4,414
Transform & Normalize	(2.5x)	11,527	*4,597*	(1.0x)	7,990	*7,990*	(4.7x)	20,491	*4,338*	(27.4x)	81,985	*2,998*	(1.0x)	14,455	*14,455*	(7.3x)	27,290	6,876
Arithmetic Mean	(4.5x)	10,240,997	*2,197,989*	(2.5x)	2,902,974	*1,683,709*	(5.1x)	6,138,595	*2,043,988*	(7.6x)	6,582,299	*1,689,427*	(2.5x)	7,494,325	*3,325,038*			
Geometric Mean	(3.1x)	47,328	*15,063*	(1.5x)	27,110	*18,626*	(3.1x)	52,161	*16,694*	(4.8x)	60,056	*12,416*	(1.7x)	66,825	*38,312*			

Table 5: **Average Time per Call (ns)** - C times listed in normal font, SIMD assembly times listed in *italics*, speedup shown to the left inside of (parentheses). Kernels with a grey background were not implemented in SIMD due to insufficient instruction set functionality. (**) DCT kernel originally coded in floating point, but implemented in fixed point integer for SIMD codes.

	AMD Athlon			DEC Alpha 21264			Intel Pentium III			Motorola G4			Sun UltraSPARC III		
Add Block	**90.0%**	0.1%	*58.9%*	-62.1%	38.6%	*-26.8%*	26.2%	10.9%	*44.8%*	-50.6%	3.1%	*-53.2%*	-3.5%	-52.7%	*-23.7%*
Block Match	**69.5%**	-5.8%	*44.2%*	-37.4%	42.5%	*17.8%*	18.2%	-9.0%	*17.5%*	-18.8%	13.9%	*5.2%*	-31.5%	-41.5%	*-84.7%*
Clip Test & Project	**8.3%**	18.1%	*35.1%*	-33.1%	4.6%	*-22.3%*	-16.1%	11.9%	*9.8%*	74.0%	1.1%	*51.2%*	-33.1%	-35.7%	*-73.9%*
Color Space	**97.6%**	-88.0%	*11.1%*	-64.1%	65.5%	*10.3%*	16.4%	7.4%	*25.7%*	-6.3%	10.6%	*10.9%*	-43.6%	4.5%	*-58.1%*
DCT**	**-25.5%**	12.5%	*4.9%*	-13.1%	55.2%	*58.2%*	37.0%	-37.3%	*18.8%*	61.0%	-19.3%	*40.0%*	-59.4%	-11.2%	*-121.8%*
FFT	**28.4%**	8.2%	*28.5%*	-35.1%	50.7%	*24.1%*	12.7%	-7.6%	*4.6%*	29.1%	-69.9%	*-31.7%*	-35.1%	18.6%	*-25.4%*
IDCT	**41.1%**	-5.9%	*35.9%*	-61.1%	54.9%	*7.1%*	8.6%	2.1%	*23.0%*	52.7%	-17.5%	*34.3%*	-41.3%	-33.6%	*-94.4%*
Max Value	**23.6%**	26.4%	*74.3%*	-69.6%	65.1%	*50.4%*	17.7%	22.5%	*71.6%*	106.3%	-25.6%	*73.7%*	-78.0%	-88.4%	*-270.1%*
Mix Stereo	**19.6%**	39.5%	*53.3%*	-57.4%	61.0%	*15.6%*	-14.1%	32.6%	*27.5%*	63.7%	-8.3%	*38.9%*	-11.8%	-124.8%	*-135.3%*
Quantize	**10.2%**	-10.2%	*9.4%*	-40.3%	37.1%	*4.5%*	47.6%	-19.4%	*26.7%*	22.8%	5.1%	*30.0%*	-40.3%	-12.5%	*-70.6%*
Short Term Analysis Filter	**-40.3%**	26.2%	*2.6%*	-61.4%	46.3%	*-9.6%*	-43.3%	22.0%	*-8.4%*	150.3%	-16.8%	*63.3%*	-5.3%	-77.7%	*-47.9%*
Short Term Synthesis Filter	**-30.5%**	30.4%	*27.6%*	-76.4%	38.8%	*-87.3%*	78.7%	-38.8%	*43.8%*	52.6%	24.5%	*64.2%*	-24.4%	-54.9%	*-48.3%*
Subsample Horizontal	**12.4%**	-4.9%	*9.6%*	-20.7%	39.5%	*26.1%*	7.4%	-3.6%	*6.6%*	28.5%	-14.7%	*13.5%*	-27.6%	-16.3%	*-55.7%*
Subsample Vertical	**-44.2%**	2.9%	*-38.9%*	-22.8%	44.1%	*42.3%*	-48.4%	16.4%	*-29.3%*	125.5%	-11.7%	*60.5%*	-10.1%	-51.7%	*-34.6%*
Synthesis Filter	**68.3%**	-12.0%	*41.4%*	-40.5%	24.9%	*-11.0%*	5.3%	-28.0%	*-6.9%*	7.3%	24.6%	*38.2%*	-40.5%	-9.4%	*-61.7%*
Transform & Normalize	**-65.7%**	57.8%	*33.1%*	-86.3%	70.7%	*-16.2%*	-35.4%	24.9%	*36.9%*	273.8%	-200.4%	*56.4%*	-86.3%	47.0%	*-110.2%*
Arithmetic Average	**16.4%**	6.0%	*26.9%*	-48.8%	46.2%	*4.8%*	7.4%	0.4%	*19.5%*	60.7%	-18.8%	*31.0%*	-35.7%	-33.8%	*-82.3%*

Table 6: **Percent Deviation from the Mean (%DM)** - for data in Table 5, which is defined as $100 \cdot \left(\frac{\bar{t}-t_i}{\bar{t}}\right)$, where t_i is the time taken on platform i, and $\bar{t}$ is the average performance (time) across all of the platforms examined for a particular kernel.

sufficient SIMD instruction set functionality to implement a SIMD version of the DCT. A C fixed-point integer DCT was substituted, which is why a 19.7x speedup is shown in Table 5, even though we did not recode it.

AMD Athlon, Intel Pentium III The AMD Athlon and Intel Pentium III processors in our study at first glance appear to be very similar; both run at 500 MHz, and, as we noted in Section 4, both share Intel's MMX SIMD integer extension. In fact, because we implemented the Intel kernel set first, it was possible to simply reuse the same code for the AMD SIMD version of many of the integer kernels in the BMKL. The SIMD integer kernels include all but the clip test, FFT, quantize, synthesis filter and transform kernels. It is interesting to observe the differences in performance between the two processors on what is often identical code. A few salient architectural differences to note [Stil99]:

- Athlon has a 64 KB instruction, 64 KB data cache, while Pentium III's L1 caches are 16 KB/16 KB respectively
- Athlon has three instruction decoders working in parallel, while Pentium III has only two
- Athlon's pipeline is 10-cycles long, while Pentium III's is 12-17 cycles; the cost of branches in the Pentium is exacerbated by Pentium III's weaker branch prediction unit
- Intel's MMX instruction latency is lower (1 cycle SIMD add/sub, 1 cycle shuffle, 3 cycle SIMD mult) compared to Athlon (2 cycle SIMD add/sub, 2 cycle shuffle, 3 cycle SIMD mult)

So, although the SIMD integer instruction set is the same in both cases, AMD's Athlon SIMD integer is a more powerful implementation; its only deficiency is the higher simple SIMD integer instruction latency when compared to the Pentium III. We can see this very clearly in Table 5, although where each architectural difference comes into play depends on the kernel.

One interesting case is that of the DCT and IDCT kernels - the Pentium III is faster on the DCT, while the Athlon is faster on the IDCT. This is surprising given that the two kernels are algorithmically quite similar - each is basically a mirror image of the other. Upon further investigation, we found that the DCT code has shuffle instructions (`pshufw`) in several places that are not scheduled well for the Athlon instruction's higher latency; a circumstance which was not duplicated in the IDCT code.

In terms of multimedia, the greatest difference between the two processors are their SIMD floating point extensions. AMD's 3DNow! is quite similar to MMX, reusing the eight x87 floating point registers as 64-bit wide SIMD registers. Intel's SSE has new 128-bit wide registers and instructions, although in the Pentium III implementation, each 128-bit instruction is actually decoded into two 64-bit wide micro-ops. This does, however, give the Pentium III more overall register space to work with. Other important architectural features of the two SIMD floating point instruction sets:

- AMD's 3DNow! FP latency (4 cycle add/sub, 4 cycle multiply) is lower than Intel's SSE (4 cycle add/sub, 5 cycle multiply, throughput of 64-bits per cycle like 3DNow!)

In the arena of floating point operations, the AMD Athlon architecture eliminates the only deficiency it had compared to the Pentium III in SIMD integer - namely, slightly higher instruction latency. From Table 6, we can see that overall SIMD AMD Athlon code does 26.9% better than average, while the Pentium III is 19.7% better than the average of all of the multimedia instruction sets. Both chips perform well, but the AMD Athlon matches or outperforms the Intel chip on all but one SIMD floating point kernel: quantize.

Quantization is a process by which a real value is converted to a discrete integer representation. An array of real double precision floating point values, $xr[]$, is converted to an array of 32-bit integers, $xi[]$, according to the function $xi[i] = \sqrt{\sqrt{xr[i]*} + 0.4054}$. Note that original C implementation utilizes a lookup table for some values (not shown in Algorithm 20).

Algorithm 20 Quantize

```
static INT32 lutab[10000];
INT32 l_end; FLOAT64 xr[]; INT32 ix[]; FLOAT64 *istep_p;
FLOAT32 temp;
for (i=0;i<l_end;i++) {
  temp=(*istep)*fabs(xr[i]);
  ix[i] = (int)( sqrt(sqrt(temp)*temp) + 0.4054);
}
```

Although SIMD floating point instruction sets include approximations for $\frac{1}{\sqrt{a}}$, rather than $\sqrt{a}$, the equivalence $\sqrt{a} = a \cdot \frac{1}{\sqrt{a}}$ can be used. The reason for AMD's lackluster performance on the quantize kernel is clear based on our earlier discussion - AMD's reciprocal square root instructions are actually scalar - they only produce one result value, based on the lower packed element. This costs 3DNow! performance for this highly square-root intensive kernel.

DEC Alpha 21264 From Tables 5 and 6 we see that the DEC Alpha platform far outstrips the other systems in terms of compiled C performance. It has been claimed that a broader multimedia instruction set would not be useful on Alpha, as an extension like Intel's MMX only fixes x86 legacy-architecture performance deficiencies which are not present in the Alpha architecture [Rubi96]. Our performance comparison makes this sound rather dubious, as the kernels programmed with the extensions from AMD, Intel and Motorola were able to not only match the performance of those on the Alpha 21264, but often exceeded it.

Motorola G4 Motorola's AltiVec was the only multimedia extension which was architected from the ground up - all of the others in some way leverage existing resources. AltiVec in many ways agrees with our design suggestions (e.g. a large number of 128-bit wide registers), although in fact the instructions included in AltiVec are far more general than those required by multimedia applications. Almost every possible operation and data type is supported, which should allow it to be applied to other application domains as well. A 128-bit register width combined with latencies that are at least as good as those found on the other (64-bit) architectures, allows AltiVec to come in with the best overall performance (31% better than average on SIMD accelerated code, according to Table 6). However, performance on a few of the kernels is still poor, especially on add block and the FFT. The add block kernel's problem has already been discussed - the 128-bit register width is actually too long for this kernel, causing unnecessary data to be loaded from memory. The reason for the poor FFT performance is not entirely clear, although a review of our code revealed that the way in which we coded data to be stored to unaligned memory could be improved.

Sun UltraSPARC IIi The performance of the VIS multimedia extension was mediocre at best, although we should note that the UltraSPARC IIi processor examined is the only one in our study running at 360 MHz (the other processors all have a 500 MHz clock). It is also the oldest multimedia instruction set we have looked at, the second to be released after HP's MAX-1 (1996). As we have pointed out, this instruction set suffers from some odd instruction choices (e.g. multiplication primitives), missing functionality (no partitioned shift operations) and a highly utilized control register that creates a bottleneck (the graphics status register).

7 New Directions

In this section we describe two new ideas for future multimedia extensions based on features we found lacking during our coding experience.

7.1 Strided Memory Access

Consider the difference between how memory is loaded into a SIMD register in the horizontal and vertical subsampling kernels. The original C sources for the horizontal and vertical subsampling kernels are in Algorithms 21 and 22 respectively. In the horizontal subsampling case, a vector load can only directly retrieve the data from one iteration of the loop into a register. For vertical subsampling, the n^{th} element from each of M loop iterations is loaded (M is the number of packed element in a register) into a register without any data rearrangement. Because SIMD applies the same operation to all of the elements of a vector, the vertical subsampling kernel can be computed more efficiently.

Algorithm 21 Subsample Horizontal

```
UINT8 *src; UINT8 *dst; INT32 width; INT32 height;
 for (j=0; j<height; j++) {
   for (i=0; i<width; i+=2) {
     ltmp = (22*(src[i-5]   + src[i+5])-52*(src[i-3]  + src[i+3])
            +159*(src[i-1] + src[i-1])+256*src[i]    + 256)>>9;
     /* clip result to UINT8 range 0..255 */
     dst[i>>1] = ltmp>255 ?  255 :  (ltmp<0 ?  0 :  ltmp);
   }
   src+= width; dst+= width>>1;
 }
```

Algorithm 22 Subsample Vertical

```
 UINT8 *src; UINT8 *dst; INT32 width; INT32 height;
 INT32 w, ltmp;
 w = width>>1;
 for (i=0; i<w; i++) {
  for (j=0; j<height; j+=2) {
   /* FIR filter with 0.5 sample interval phase shift */
   ltmp = (228*(src[w*j] + src[w*j+1])+70*(src[w*(j-1)] + src[w*j+2])
     -37*(src[w*j-2] + src[w*j+3])-21*(src[w*j-3] + src[w*j+4])
     +11*(src[w*j-4] + src[w*j+5])+5*(src[w*j-5] +
src[w*j+6])+256)>>9;
   /* clip result to UINT8 range 0..255 */
   dst[w*(j>>1)]=ltmp>255 ?  255 :  (ltmp<0 ?  0 :  ltmp);
  }
  src++;
  dst++;
 }
```

With current SIMD architectures, when registers contain the data for a single loop iteration, either operations on some of the packed elements must be nullified, or significant overhead must go into transposing the data, wasting computation. Although this worked acceptably well in the DCT and IDCT kernels, there are some cases, such as image processing, when it is not feasible to transpose an image - the overhead is far too great.

We propose that SIMD architectures implement strided load and store instructions to make the gathering of non-adjacent data elements more efficient. This is similar to the prefetch mechanism in AltiVec, except that the data elements would be assembled together by the hardware into a single register, rather than simply loaded into the cache. Of course such a memory operation would necessarily be slower than a traditional one, but it would cut down immensely on the overhead that would have to go into reorganizing data as loaded from memory. Strided loads and stores would have three operands:

Instruction	Syntax
Load Strided	`lvxstrd vD, rA, rB`
Store Strided	`stvstrd vD, rA, rB`

where in each case `rA` is the base address and `rB` contains a description of the memory access pattern:

width — width of the data elements to be loaded [2 bits], 00 = 8-bits, 01 = 16-bits, 10 = 32-bits, 11= 64-bits

offset — number of bytes offset from the base address from which to begin loading [6 bits], interpreted as an signed value: -32..+31

stride — number of bytes between the effective address of one element in the sequence and the next [24 bits], interpreted as a signed value: -8388608..+8388607

The color space conversion kernel is an excellent example of where strided load and store instructions could be used. Pixel data consists of one or more channels or bands, with each channel representing some independent value associated with a given pixel's (x, y) position. A single channel, for example, represents greyscale, while a three (or more) channel image is typically color. The band data may be interleaved (each pixel's red, green, and blue data are adjacent in memory) or separated (e.g. the red data for adjacent pixels are adjacent in memory). In image processing algorithms such as color space conversion we operate on each channel in a different way, so band separated format is the most convenient for SIMD processing. Converting from the RGB to YC_BC_R color space is done through the conversion coefficients shown in Algorithm 23.

Algorithm 23 Color Space Conversion

```
UINT8 *rowp, *y_p, *u_p, v_p;
INT32 red = *rowp++, green = *rowp++, blue = *rowp++;
*y_p++ = +0.2558*red + 0.5022*green + 0.0975*blue + 16.5;
*u_p++ = -0.1476*red - 0.2899*green + 0.4375*blue + 128.5;
*v_p++ = +0.4375*red - 0.3664*green - 0.0711*blue + 128.5;
```

Algorithm 24 replaces thirty-eight instructions in the original AltiVec color space conversion kernel (the corresponding code fragment is listed in [Sling00d]). In the original AltiVec code, it was necessary to load six permute control vectors (each 128-bits wide) before executing the six `vperm` instructions required to rearrange the data into band separated format.

Algorithm 24 Modified Color Space Conversion

```
rgb_to_yuv:
oris    r11,r11 RED_PATTERN
oris    r12,r12,GREEN_PATTERN
oris    r13,r13,BLUE_PATTERN
lvxstrd v28,0,r3,r11
;; v28: |r0|r1|r2|r3|r4|r5|r6|r7|r8|r9|rA|rB|rC|rD|rE|rF|
lvxstrd v29,0,r3,r12
;; v29: |g0|g1|g2|g3|g4|g5|g6|g7|g8|g9|gA|gB|gC|gD|gE|gF|
lvxstrd v30,0,r3,r13
;; v30: |b0|b1|b2|b3|b4|b5|b6|b7|b8|b9|bA|bB|bC|bD|bE|bF|
```

Traditional SIMD data communication operations have trouble with data which is not aligned on boundaries which are powers of two - in the case of color space conversion, visually adjacent pixels from each band are spaced 3 bytes apart. Strided loads and stores are by definition unaligned, so this would need to be handled by the load/store hardware in the CPU. It would also make sense to have additional versions of these instructions which would be a hint to circumvent the cache (on a load) or to not do write-allocation (on a store) if the cache lines containing the strided data elements would not be of near-term utility.

7.2 Superwide Registers

Generally, multimedia data is stored in a packed format and is loaded into registers into the same format. Frequently, unpacking is required before operations are performed, and the unpacked data, of course, no longer fits within a single register. We therefore propose:

- registers that are wider than the data loaded into them
- implicit unpack with load
- implicit pack with store

Our design is in some ways similar to the as yet unimplemented MIPS MDMX instruction set [MIPS97]. The MIPS MDMX extension has a 192-bit accumulator register as its architectural cornerstone, with the more usual style of register to register SIMD operations also included; non-accumulator SIMD operations share the 64-bit floating point datapath. The destination of normal SIMD instructions can be either another SIMD register or the accumulator (to be loaded or accumulated). When accumulating packed unsigned bytes, the accumulator is partitioned into eight 24-bit unsigned slices. Packed 16-bit operations cause the accumulator to be split into four 48-bit sections. This extra width allows for multiple accumulations to occur without overflow.

What is good about the MDMX accumulator approach is that implicit width promotion provides an elegant solution to overflow and other issues caused by packing data as tightly as possible. For example, multiplication is semantically difficult to deal with on SIMD architectures because the result of a multiply is longer than either operand [Lee97c]. This problem is avoided by the MDMX accumulator, because there is enough extra space to prevent overflow.

Fixed point arithmetic is also more precise because full precision results can be accumulated, and the total rounded only once at the end of the loop. Similarly, most scalar multimedia algorithms only specify saturation at the end of computation. This is because it is more precise to saturate once rather than at every step of the algorithm. For example, if we are adding three signed 16-bit values:

Saturation at Every Step: $32760 + 50 - 20 = 32747$

Saturation at Last Step: $32760 + 50 - 20 = 32767$

Unfortunately, in SIMD architectures where the packed elements maximally fill the available register space, the choice is either to be imprecise (compute with saturation at every step) or loose parallelism (explicitly promote the input data to be wider). Saturating arithmetic can also produce unexpected results since, unlike normal addition, the order of operations matters.

While the MIPS MDMX solution may seem elegant in principle, it ignores the architectural side of actually making such a design fast. The accumulator is a singular (unique) shared resource, and as such has a tendency to limit instruction level parallelism. We found that the existence of only a single accumulator was a severe handicap to avoiding data dependencies. On a non-accumulator but otherwise superscalar architecture it is usually possible to perform some other useful, non data dependent operation in parallel so that the processing can proceed at the greatest degree of instruction level parallelism possible. On MDMX all computations which need to use the accumulator must proceed serially.

Low Precision		High Precision	
Storage	**Computation**	**Storage**	**Computation**
8U	12Sx16	8U	24Sx8
16S	24Sx8	16S	48Sx4
32S	48Sx4	-	-
32FP	32FPx4	-	-

Table 7: **Supported Data Types** - data types are divided into storage (how it is stored in memory) and computation (the width of arithmetic and other instruction elements)

Supported Data Types To avoid the problems with MDMX, we suggest that a normal register file architecture be used, with the entire SIMD register file made wide (for example, 192-bits). Supported memory (load and store) data types include those that we have seen to be of importance in multimedia for intermediate and storage formats: 8-bit unsigned, 16-bit signed, 32-bit signed and single precision floating point. The data types actually supported by packed arithmetic operations are different: 12-bit signed, 24-bit signed, 48-bit signed and single precision floating point. Depending on the algorithm it may be desirable to utilize the extra bits of a superwide register for either accumulation or data parallelism. We suggest supporting two memory widths: 64-bits (high-precision) and 128-bits (low-precision), which are unpacked to different computational widths. This also allows for better matching with algorithms that have different natural widths (e.g. add block, which works best with a 64-bit vector length).

Packed floating point data types present an interesting design choice - we can either operate on:

1. four single precision values in parallel (not using 16 of the bits in each register element)
2. four single precision values which have been expanded to a 48-bit extended precision format
3. six single precision values, exactly filling a 192-bit register

The downside of the second solution is that SIMD results may not exactly match scalar results. In addition, the latency of many floating point operations depends heavily on the precision being computed, so a more precise operation is a higher latency one. The third solution, although attractive for its additional data parallelism, is problematic because it would require its own set of data rearrangement instructions based on a six element (rather than four or eight element) vector. For our sample design, we chose the first option.

Supported Operations Because we have fundamentally changed the treatment of multimedia data types, we also need to reexamine which operations are still valuable in this new light. Several instructions are no longer useful:

- average - the only useful average instruction data type within our workload was for 8-bit unsigned values. Average is only really useful on existing multimedia architectures because it allows for computation without width

promotion. On a superwide register architecture average instructions are unnecessary, since the requisite functionality can be synthesized through shift and add (operations which are useful in and of themselves), and there is already sufficient precision.

- saturating arithmetic - saturation is done implicitly during packing. Of course, max and min instructions can always be used if an exotic type of clamping (e.g. 9-bit signed in the IDCT) need to be supported.
- pack/unpack - performed implicitly with loads and stores
- truncating multiplication - *truncation* predefines a set number of result bits to be thrown away. This primarily has application when multiplying n-bit fixed point values with fractional components which together take up a total of n-bits of precision. Unfortunately, this plays havoc with precision since it is usually desirable to truncate once, at the end of a fixed-point computation, rather than at every step. Because all of the high precision computational data types in our design are more than wide enough to hold the product of two storage data types, overflow is never a problem.

In [Sling00d] we present our proposed instruction set (97 instructions in total) for a superwide register SIMD multimedia extension. Unlike existing instruction sets which are fundamentally byte-based, this instruction set is centered around quantities which are multiples of 12-bits wide. This can be though of as four extra bits of precision for every byte of actual storage data loaded.

Example A small example of how to code for the proposed architecture is shown in Figure 1. Note that load and store operations specify both the computational and storage data type.

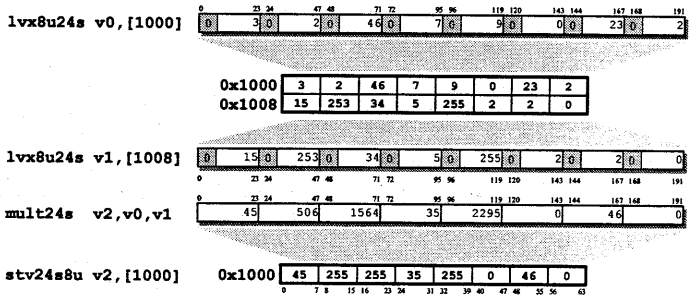

Figure 1: **Superwide Registers Example**

8 Summary

8.1 Useful Features

Many of the architectural features of existing multimedia instruction sets attempt to get around the limitations of tightly packed SIMD registers. With a superwide register architecture, many of the reasons for these features are eliminated, creating a simpler overall design. Our summary distills our conclusions about standard tightly packed SIMD, although we note were there are differences introduced by a superwide register approach.

8.1.1 Register File

- Sharing the floating point datapath is usually preferable to the integer data path because there is no contention with pointer and loop variables, and less chance of affecting the critical path of the processor. If no existing data path is to be shared by SIMD instructions, a 128-bit wide data path is optimal for most multimedia algorithms, 192-bits in the case of a superwide architecture.
- Multimedia algorithms can take advantage of large register files - we suggest *at least* 16 128-bit registers, or 32 64-bit registers.

8.1.2 Data Types

- 8-bit signed data types are not useful. 8-bit unsigned data types are most often used for storage, rather than computation.
- 16-bit signed data types are the most common; they are the intermediate (computational) data type for video and the storage (and sometimes computational) data type for audio and speech algorithms. Unsigned 16-bit values are not useful.
- 32-bit signed values are most often used for accumulation. Unsigned 32-bit values are not useful.
- Single precision floating point (32-bit) is found in the audio and 3D graphics (geometry) kernels. We did not come across a multimedia algorithm which required double precision (64-bit) floating point.

8.1.3 Integer Arithmetic

- Saturation prevents overflow in a fast, numerically acceptable way for SIMD operations, although with our proposed superwide register architecture, saturation is really not needed except when packing.
- Max and min operations are an efficient way to perform SIMD comparisons as well as clamping to arbitrary ranges. They are useful at all computational data widths.
- Average instructions we found only to be useful in the MPEG encode block match kernel for interpolation (8-bit unsigned data type). They are less useful on a superwide register architecture, because there averaging can be done through an add and subsequent shift right by one bit without unpacking to a wider width.

- Shift operations of all types are useful at all data widths - they are critical for fixed point arithmetic, and also provide an efficient means for data realignment and division and multiplication by powers of two.

- Sum of absolute difference instructions are only useful for MPEG encoding and other video encoding algorithms which utilize motion compensation (block match kernel - 8-bit unsigned data type).

8.1.4 Floating Point Arithmetic

- $\frac{1}{\sqrt{x}}$approximation instructions are useful; through multiplication they can also estimate $\sqrt{x}$ and $\frac{1}{x}$. A full precision version of this instruction is not necessary, as the Newton-Raphson method can always be used to improve the precision of the approximation.

- Exceptions and sophisticated rounding modes (as specified by the IEEE floating point standard) are not necessary for multimedia; in any instance of where these might be used it is possible to substitute a reasonable value that will allow computation to continue unhindered, and still produce an acceptable result.

8.1.5 Data Rearrangement

- A full permute operation (as is found in AltiVec) is very flexible, but is probably overkill for most multimedia applications where data rearrangement patterns can be handled by simpler data rearrangement operations. However, it should be noted that in AltiVec the `vperm` instruction serves double duty as a means for aligning unaligned data loads, so its capabilities are basically free.

- [Lee00] presents a novel set of simple data communication primitives which can perform all 24 permutations of a 2x2 matrix in a single cycle on a processor with dual data communication functional units. We endorse this technique because any larger data communication problem can be decomposed into 2x2 matrices, and because most multimedia data rearrangement patters are simple; they can be done in a single cycle. [Lee00]'s instructions are preferable to `vperm` because they do not require a permutation control vector to first be loaded into memory, as their data communication patterns are statically defined.

8.1.6 Memory Operations

- Hardware support to efficiently handle memory accesses that are not aligned are expensive in both area and timing [Thak99]. Ideally, data would always be aligned by software (e.g. the compiler or run-time architecture). In some situations it is impossible to guarantee alignment. The strided load and store operations which we have proposed would be in many cases inherently unaligned, making hardware support a requirement. Also, for example, in the motion compensation step of MPEG video coding, unaligned memory access is needed depending on the motion vector [Kuro98], as the addresses of the reference macroblock can be random depending on the type of motion search being performed.

- Allowing only aligned memory accesses (and synthesizing unaligned accesses in software) can potentially perform better than unaligned access implemented in hardware. However, silently accepting an unaligned address and forcing it to be aligned (as in AltiVec) is a bad idea as it can allow alignment errors (typically very difficult to track down because they are intermittent) to go unnoticed. Instead, an exception should be raised when an unaligned access occurs, or hardware should support unaligned memory access directly.

8.2 Bottlenecks and Unnecessary Features

- Instruction primitives (such as the multiplication instruction primitives found in Sun's VIS) are a bad idea, as they decrease instruction decoding bandwidth, increase register pressure, and are not useful in and of themselves. Even if the atomic version of an operation may be slow, it is preferable because it is much easier to upgrade an instruction's latency in the next revision of an architecture than it is to implement entirely new instructions, rendering the previous instructions and any related ones useless.

- Motorola's AltiVec extension includes pixel pack and unpack instructions for converting between 32-bit true color and 16-bit color representations which we did not find useful in the BMKL. Similarly, AltiVec includes approximations for log_2 and exp_2, which also went without application in our workload; they are used in lighting algorithms for 3D rendering.

- Sun's VIS includes edge instructions for dealing with boundaries in 2D image processing, as well as array instructions for volumetric imaging. Neither type of instruction was found to be useful to the Berkeley multimedia workload.

- In general, a singular (unique) resource (such as a control register, or accumulator) is a potential bottleneck if it will be highly utilized. In the case of Sun's VIS graphics status register (GSR), their bottleneck could have been avoided if SIMD shift instructions and better data communication primitives had been included. As it was, the GSR ended up being used to synthesize this missing functionality, beyond its original designed purpose. The MIPS MDMX accumulator register which we briefly discussed is another example of this type of problem.

8.3 New Directions

In addition to analyzing how well current multimedia instruction set features map to multimedia workloads, we also pro-

posed two new directions for multimedia instruction sets.

- Because SIMD architectures apply the same operation to all of the elements in a packed register, there are many cases where data is not optimally organized as loaded from memory. This occurs when working with 2D data types, such as video frames; either row or column processing will not be natively arranged in a way that is amenable to SIMD style processing. Typically, we would like to load M data elements into a vector register, with each element being loaded starting at some base address and separated from each other by a constant byte offset. Similar to scatter-gather operations from traditional vector architectures, we proposed implementing strided loads and stores for packed registers. These are specified in a way that is similar to how prefetch streams are specified in AltiVec.
- Based on the observation that storage and computational data types are almost always different, we proposed a superwide register architecture, which eliminates much of the explicit packing and unpacking overhead that typically makes SIMD processing progress at less than its maximal degree of data parallelism. We found that this fundamental change in how data types are handled had significant implications for instruction set design.

References

[Allen99] Gregory E. Allen, Brian L. Evans, Lizy K. John, "Real-Time High-Throughput Sonar Beamforming Kernels Using Native Signal Processing and Memory Latency Hiding Techniques," *Proc. of the 33rd IEEE Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, California, October 24-27, 1999, pp. 137-141

[AMDOpt] Advanced Micro Devices, *AMD Athlon Processor x86 Code Optimization Guide*, Publication 22007G/0, April 2000, *http://www.amd.com/products/cpg/athlon/techdocs/index.html*, retrieved April 24, 2000

[AMD00] Advanced Micro Devices, Inc., "AMD Athlon Processor x86 Code Optimization Guide," Publication #22007, Rev. G, April 2000, *http://www.amd.com/products/cpg/athlon/techdocs/pdf/22007.pdf*, retrieved April 24, 2000

[AMD99] Advanced Micro Devices, Inc., "AMD Athlon Processor Technical Brief," Publication #22054, Rev. D, December 1999, *http://www.amd.com/products/cpg/athlon/techdocs/pdf/22054.pdf*, retrieved April 24, 2000

[AMDWP] Advanced Micro Devices, "3DNow! Technology vs. KNI," White Paper, *http://www.amd.com/products/cpg/3dnow/vskni.html*, retrieved April 24, 2000

[Bhar98] R. Bhargava, L. K. John, B. L. Evans, R. Radhakrishnan, "Evaluating MMX Technology Using DSP and Multimedia Applications," *Proc. of the 31st IEEE Intl. Symp. on Microarchitecture (MICRO-31)*, Dallas, Texas, November 30-December 2, 1998, pp. 37-46

[Burd] Tom Burd, "CPU Info Center: General Processor Info," *http://bwrc.eecs.berkeley.edu/CIC/summary/local/summary.pdf*, retrieved April 24, 2000

[Carl97] David A. Carlson, Ruben W. Castelino, Robert O. Mueller, "Multimedia Extensions for a 550-MHz RISC Microprocessor," *IEEE Journal of Solid-State Circuits*, Vol. 32, No. 11, November 1997, pp. 1618-1624

[Chen96] Wilam Chen, H. John Reekie, Sunil Bhave, Edward A. Lee, "Native Signal Processing on the UltraSparc in the Ptolemy Environment," *Proc. of the 30th Annual Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, California, November 3-6, 1996, Vol. 2, pp. 1368-1372

[Comp00] Compaq Computer Corporation, "Alpha 21264 Microprocessor Hardware Reference Manual," Part No. DS-0027A-TE, February 2000, *http://www.support.compaq.com/alpha-tools/documentation/current/21264_EV67/ds-0027a-te_21264_hrm.pdf*, retrieved April 24, 2000

[Hans96] Craig Hansen, "MicroUnity's MediaProcessor Architecture," *IEEE Micro*, Vol. 16, No. 4, August 1996, pp. 34-41

[IAOpt] Intel Corporation, *Intel Architecture Optimization Reference Manual*, *http://developer.intel.com/design/pentiumii/manuals*, retrieved April 24, 2000

[IEEE91] IEEE, "IEEE Standard Specifications for the Implementation of 8x8 Inverse Discrete Cosine Transform", *IEEE Standard* 1180-1990, M. T. Sun ed., IEEE, 1991

[Intel97] Intel Corporation, "Intel Introduces The Pentium Processor With MMX Technology," January 8, 1997 Press Release, *http://www.intel.com/pressroom/archive/releases/dp010897.htm*, retrieved April 24, 2000

[Intel99a] Intel Corporation, "Intel Architecture Software Developer's Manual, Volume 1: Basic Architecture," Publication 243190, 1999, *http://developer.intel.com/design/pentiumii/manuals/24319002.PDF*, retrieved April 24, 2000

[Intel00a] Intel Corporation, "IA32 Intel Architecture Software Developer's Manual with Preliminary Intel Pentium 4 Processor Information Volume 1: Basic Architecture," *http://developer.intel.com/design/processor/future/manuals/24547001.pdf*, retrieved September 26, 2000

[Intel00b] Intel Corporation, "Intel Announces New NetBurst Micro-Architecture for Pentium IV Processor," *http://www.intel.com/pressroom/archive/releases/dp082200.htm*, retrieved September 27, 2000

[Kesh99] Jagannath Keshava, Vladimir Pentkovski, "Pentium III Processor Implementation Tradeoffs," Intel Technology Journal, Quarter 2, 1999, *http://developer.intel.com/technology/itj/q21999/pdf/impliment.pdf*, retrieved April 24, 2000

[Kien99] Tim Kientzle, "Implementing Fast DCTs," *Dr. Dobb's Journal*, Vol. 24, No. 3, March 1999, pp. 115-119

[Kohn95] L. Kohn, G. Maturana, M. Tremblay, A. Prabhu, G. Zyner, "The Visual Instruction Set (VIS) in UltraSPARC," *Proc. of Compcon '95*, San Francisco, California, March 5-9, 1995, pp. 462-469

[Kuro98] Ichiro Kuroda, Takao Nishitani, "Multimedia Processors," *Proc. of the IEEE*, Vol. 86 No. 6, June 1998, pp. 1203-1221

[Lawl92] Patricia K. Lawlis, "Ada Outperforms Assembly: A Case Study," *Proceedings of TRI-Ada '92*, Orlando, Florida, November 16-20, 1992, *http://www.acm.org/sigs/sigada/education/pages/lawlis.html*, retrieved October 20, 2000

[Lee00] Ruby B. Lee, "Subword Permutation Instructions for Two-Dimensional Multimedia Processing in MicroSIMD Architectures," *Proc. of IEEE Intl. Conf. on Application-Specific Systems, Architectures, and Processors*, July 10-12, 2000, Boston, Massachusetts, pg. 3-14

[Lee97c] Ruby B. Lee, "Multimedia Extensions for General Purpose Processors," *Proc. of the IEEE Workshop on VLSI Signal Processing*, Leicester, UK, November 3-5, 1997, pp. 1-15

[MIPS97] MIPS Technologies, Inc., "MIPS Extension for Digital Media with 3D," WhitePaper, March 12, 1997 *http://www.mips.com/Documentation/ isa5_tech_brf.pdf*, retrieved April 24, 2000

[Moto00] Motorola Inc., "MPC7400 RISC Microprocessor User's Manual, Rev. 0," Document MPC7400UM/D, March 27, 2000, *http://www.mot.com/SPS/PowerPC/teksupport /teklibrary/manuals/MPC7400UM.pdf*, retrieved April 24, 2000

[Naka96] Jill Nakashima, Ken Tallman, "The VIS Advantage: Benchmark Results Chart VIS Performance," White Paper, October 1996, *http://www.sun.com/microelectronics/vis/*, retrieved April 24, 2000

[Nguy99] Huy Nguyen, Lizy Kurian John, "Exploiting SIMD Parallelism in DSP and Multimedia Algorithms Using the AltiVec Technology," *Proc. of the 1999 Interational Conf. on Supercomputing*, Rhodes, Greece, June 20-25, 1999, pp. 11-20

[Noer] Kim Noer, "Heat Dissipation Per Square Millimeter Die Size Specifications," *http://home.worldonline.dk/~noer/*, retrieved April 24, 2000

[Norm98] Kevin B. Normoyle, Michael A. Csoppenszky, Allan Tzeng, Timothy P. Johnson, Christopher D. Furman, Jamshid Mostoufi, "UltraSPARC-IIi: Expanding the Boundaries of a System on a Chip," *IEEE Micro*, Vol. 18, No. 2, March/April 1998, pp. 14-24

[Patt96] David A. Patterson, John L. Hennessy, *Computer Architecture: A Quantitative Approach*, Second Edition, 1996 Morgan Kaufman Publishers, Inc.

[Rang99] Parthasarathy Ranganathan, Sarita Adve, Norman P. Jouppi, "Performance of Image and Video Processing with General-Purpose Processors and Media ISA Extensions," *Proc. of the 26th Annual Intl. Symp. on Computer Architecture*, May 2-4, 1999, Atlanta, Georgia, pp. 124-135

[Rath96] Selliah Rathnam, Gert Slavenberg, "An Architectural Overview of the Programmable Multimedia Processor, TM-1," *Proc. of COMPCON '96*, February 25-28, 1996, Santa Clara, California, pp. 319-326

[Rice96] Daniel S. Rice, "High-Performance Image Processing Using Special-Purpose CPU Instructions: The UltraSPARC Visual Instruction Set," University of California at Berkeley, Master's Report, March 19, 1996

[Rubi96] Paul Rubinfeld, Bob Rose, Michael McCallig, "Motion Video Instruction Extensions for Alpha," White Paper, October 18, 1996 *http://www.digital.com/alphaoem/papers/pmviabstract.htm*, retrieved April 24, 2000

[Sling00a] Nathan T. Slingerland, Alan Jay Smith, "Design and Characterization of the Berkeley Multimedia Workload," *University of California at Berkeley Technical Report CSD-00-1122*, December, 2000

[Sling00c] Nathan T. Slingerland, Alan Jay Smith, "Multimedia Instruction Sets for General Purpose Microprocessors: A Survey," *University of California at Berkeley Technical Report CS-00-1124*, December, 2000

[Sling00d] Nathan T. Slingerland, Alan Jay Smith, "Measuring the Performance of Multimedia Instruction Sets," *University of California at Berkeley Technical Report CSD-00-1125*, December 2000

[Stil99] Andreas Stiller, "Architecture Contest," *c't Magazine*, Vol. 16/99, *http://www.heise.de/ct/english/99/16/092/*, retrieved December 2, 2000

[Sun97] Sun Microsystems Inc., "UltraSPARC-IIi User's Manual," 1997, Part No. 805-0087-01, *http://www.sun.com/microelectronics/UltraSPARC-IIi/docs/805-0087.pdf*, retrieved April 24, 2000

[Thak99] Shreekant (Ticky) Thakkar, Tom Huff, "The Internet Streaming SIMD Extensions," *Intel Technology Journal*, Q2 1999, *http://developer.intel.com/technology/itj/q21999.htm*, retrieved April 24, 2000

[Trem96b] Marc Tremblay, J.Michael O'Connor, Venkatesh Narayanan, Liang He, "VIS Speeds New Media Processing," *IEEE Micro*, Vol. 16, No. 4, August 1996, pp. 10-20