DSP-bezogene Spracherweiterungen: Möglichkeiten und Grenzen

Markus Willems, Marek Jersak, Vojin Živojnović

Lehrstuhl für Integrierte Systeme der Signalverarbeitung
RWTH Aachen

1 Einleitung

Hochsprachen-Compiler für digitale Signalprozessoren (DSPs) gehören zu den wenigen Produkten der Signalverarbeitungs-Industrie, über die eine kontroverse Debatte bezüglich ihrer Nützlichkeit geführt wird. Fakt ist, daß heute für nahezu alle angebotenen DSPs Hochsprachen-Compiler zur Verfügung stehen. Unterstützt wird vor allem C, es existieren auch Compiler für C^{++} [1], Ada und Pascal [2]. Fakt ist aber auch, daß signifikante Anteile der DSP-Programmierung in Assembly erfolgen. Damit ergibt sich die provozierende Frage:

DSP-Compiler: Marketing-Tool, oder zukünftig auch Produktions-Tool?

Die Motivationen für eine Programmierung in einer Hochsprache sind allgemein bekannt:

- Produktivität
  die höhere Abstraktion des Hochsprache macht das Programmieren komfortabler und erlaubt eine schneller Beschreibung der Applikation. Insbesondere Änderungen und Umstrukturierungen sind wesentlich einfacher durchzuführen als auf der Assembly-Ebene.
- Testen und Debuggen
  die Hochsprache erlaubt effizientes Testen und Debuggen
- Portierbarkeit
  der Code kann auf andere Architekturen übertragen werden
- Wiederverwendbarkeit
  getesteter Code kann wiederverwendet werden, unabhängig von der Architektur
- Wartbarkeit
  Änderungen und Erweiterungen sind wesentlich einfacher einzufügen als im Assembly-Code

Die grundsätzliche Aufgabe des Compilers ist die korrekte Übersetzung der Funktionalität aus der Hochsprache in die Maschinensprache. Darüber hinaus bestehen applikationsabhängig verschiedenen Anforderungen an den Compiler. Solange Speicherbedarf und Laufzeit keine kritischen Größen darstellen, steht die schnelle Compilierung der Funktionalität im Vordergrund, und tatsächlich sind viele in der Literatur angesprochenen Optimierungen Compilezeit-Optimierungen.

Eine Abkehr von diesem Prinzip der Compilezeit-Optimierung wird erst dann notwendig, wenn die zugrundeliegende Hardware zur Abarbeitung des generierten Codes aus Komplexitätsgründen nicht mehr in der Lage ist. Hier stellen die Signalprozessoren einen Extremfall dar: der Anwendungsbereich erfordert im allg. die schnelle Bearbeitung numerisch intensiver Aufgaben. Dabei ist aus Kosten- und Leistungsverbrauchs-Gründen ein minimaler Speicherbedarf anzustreben, da der auf dem Chip angeordnete Speicher eine knappe Resource ist und die Erweiterung um externen Speicher zusätzliche Kosten verursacht. Somit stellt sich für den Compiler die Aufgabe, extrem
- laufzeiteffizienten und
- speichereffizienten Code zu generieren.

Die Compilierzeit tritt durch diese Anforderungen in den Hintergrund.

Es ist bekannt, daß state-of-the-art Compiler bzgl. dieser beiden Anforderungen sehr oft hinter dem Notwendigen zurückbleiben. Daher gehört auch heute, trotz der Verfügbarkeit von Compilern, die Programmierung in Assembly zum Handwerk des DSP-Programmierers.

Mit DSPstone [3] wurde erstmals eine Methodik vorgetellt, die eine Erweiterung der erwähnten qualitativen Aussagen durch quantitative ermöglicht. Die Methodik basiert auf dem in Fig.1 dargestellten Prinzip.

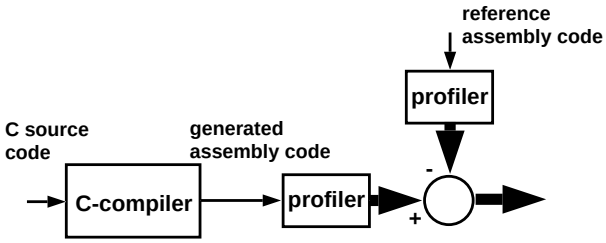

Abbildung 1: Das DSPstone-Prinzip

Für verschiedene funktionale Kerne wird der C-Quellcode compiliert und bzgl. des Speicherbedarfes und der Anzahl der notwendigen Instruktionen mit dem optimalen Assembly-Code verglichen. Diese Methodik gibt dem Anwender Auskunft über den zu erwartenden Verlust bei der Verwendung des C-Compilers, verglichen mit der optimalen Assembly-Version.

Wird mit DSPstone die Frage nach dem 'Wieviel?' beantwortet, so erlaubt der Vergleich von Compiler-generiertem und optimalem Assembly-Code auch eine Aussage zum 'Warum?'. Die Behandlung des 'Warum?' sowie Ansätze zur Verbesserung der Compiler-Effizienz sind Gegenstand der folgenden Abschnitte.

Ausgangspunkt ist in Abschnitt 2 die mikroskopische Betrachtung der Compiler-Ineffizienzen. Diese erfolgt exemplarisch anhand des FIR-Filters und stellt die vom Compiler nicht erkannten bzw. vom Compiler nicht erkennbaren Optimierungsmöglichkeiten vor. Als eine Konsequenz der Untersuchungen werden in Abschnitt 3 zwei abstrakte Ansätze zur Spracherweiterung vorgestellt: der architekturspezifische Ansatz geht von einem abstrakten Prozessormodell aus, während der anwendungsspezifische Ansatz von der gezielten Unterstützung DSP-spezifischer Applikationen und Operationen ausgeht. Abschnitt 4 beschreibt die Auswirkungen eines konkreten Spracherweiterungs-Konzepts für das im Abschnitt 2 vorgestellte Beispiel des FIR-Filters. Es zeigt sich, daß signifikante Verbesserungen der Code-Qualität erzielbar sind. Die Diskussion der Auswirkungen von Spracherweiterungen auf den Abstraktionsgrad der Hochsprache wird in Abschnitt 5 behandelt, an den sich die Zusammenfassung und der Ausblick in Abschnitt 6 anschließen.

2 Mikroskopische Analyse

Die Analyse der Unterschiede zwischen Compiler-generiertem und optimalen Code wird im folgenden exemplarisch für das FIR-Filter dargestellt. Natürlich können die Ergebnisse für dieses Beispiel nicht beliebig übertragen werden, doch die gefundenen Ineffizienzen erlauben einen konkreten Einblick in die Ineffizienzen des Compiliervorgangs. Als Beispielarchitektur dient der NEC7701x. Dieser Festkomma-Prozessor mit einer 16-bit Busstruktur bietet 2 getrennte Datenspeicher, auf die parallel zugegriffen werden kann, sowie 8 universell verwendbare, 40-bit breite Register .

Die Funktionalität des FIR-Algorithmus wird beschrieben durch

$$y(k) \quad = \quad \sum_{i=0}^{N-1} x(k-i)h(i)$$

Dabei ist $X = \{x(1), x(2), ...\}$ die Sequenz der Eingangswerte, und $H = \{h(1), ..., h(N-1)\}$ der Koeffizienten-Vektor. Zu jedem Zeitpunkt k wird ein neuer Eingangswert $x(k)$ in die Berechnung aufgenommen. Das entspricht einer Verschiebung der Datenwerte relativ zu den Filterkoeffizienten um jeweils einen Wert, und dem Herausfallen von $x(k-N)$ aus den relevanten Datenwerten.

Ein C-Code unter Verwendung der Zeiger-Adressierung lautet:

```
1 y = 0;
2 for (i=0; i < (N-1); i++){
3    y    += *ph-- * *px;
4    *px-- = *px2--;
5 }
6 y  += *ph-- * *px;
7 *px = x0;
```

Die Initialisierung der Koeffizienten und Datenwerte ist im untersuchten Code nicht mehr erafßt. Zunächst soll der optimale Assembly-Code vorgestellt werden (für N=16):

```
1   clr(r4)             r1= *dp1%%      r5= *dp5%%;
2   rep 14;
3      r4= r4+ r1h* r5h    r1= *dp1%%      r5= *dp5%%;
4   r4= r4+ r1h* r5h    r1= *dp1        r5= *dp5%%;
5   r4= r4+ r1h* r5h    r3= *dp2;
6   r4= r4 SLL r0l      *dp1= r3h;
7   *dp4= r4h;
```

Es liegt eine nahezu maximale Parallelisierung vor, neben jeder arithmetischen Operation werden eine oder zwei Ladeoperationen parallel durchgeführt. Das gleichzeitige Durchführen von zwei Ladeoperationen ist nur möglich, wenn die Operanden in unterschiedlichen Speicherbänken angeordnet sind. Dieses erlaubt im vorliegenden Beispiel insbesondere die Ausführung der inneren Schleife mit einer Instruktion. Dabei wird die `mac` Operation ausgenutzt. Das parallel stattfindende Laden benutzt die Modulo-Adressierung. Da der NEC7701x eine fractional-Arithmetik realisiert und daher die abzuspeichernden Bits im hohen Ergebnisregister erwartet, muß zur Realisierung des Integer-Ergebnisses noch eine explizite Verschiebung stattfinden. Die Anzahl der zu verschiebenden Stellen steht in einem Register, damit ergibt sich die Möglichkeit der Parallelisierung einer weiteren Lade-Operation.

Der vom Compiler generierte Assembly-Code ergibt sich zu:

```
1  CLR (r5);
2  LOOP 15 {
3    dp5 = r0l;
4    r0 = r0- 1          r1 = *dp6;
5    r2 = *dp7--;
6    r1 = r1h* r2h;
7    r1 = r1 SLL 15;
8    r5 = r5+ r1         r1 = *dp5;
9    *dp6-- = r1h;
10 }
11 r0 = *dp6;
12 r1 = *dp7;
13 r0 = r0h* r1h;
14 r0 = r0 SLL 15;
15 r5 = r5+ r0;
16 r0 = *ANON3:Y;
```

Es fällt sofort auf, daß die Parallisierung wesentlich geringer ausgefallen ist als im optimalen Code. Dieses hat mehrere Gründe:

- Daten sind nicht in verschiedenen Speicherbereichen angeordnet, damit entfällt die Möglichkeit der Parallisierung z.B der Zeilen 4 und 5. Der Compiler benutzt prinzipiell nicht den X-Speicher zur Ablage von Daten. Damit sind wesentliche Optimierungsmöglichkeiten von vorneherein ausgeschlossen. Zu dieser ineffizienten Speicherzuweisung besteht aber keine zwingende Notwendigkeit. Dem Compiler steht bereits die Information bzgl. der Konfiguration des Prozessors zur Verfügung, daher könnte er prinzipiell die Vorteilhaftigkeit alternativer Speicherzuweisungen erkennen. Hier wird deutlich, daß die DSP-Architektur durch die fehlende Orthogonalität (d.h. in diesem Fall, nicht alle Speicherplätze sind alternativ zu gleichen Kosten nutzbar) eine zusätzliche Komplexität in den Compilierungs-Vorgang einbringt, die derzeit von keinem Compiler in geeigneter Weise berücksichtigt wird.

- Die Linksverschiebung wird nicht durch den Wert in einem Register angegeben, sondern absolut. Damit ist kein paralleles Abarbeiten möglich. Eine Zuweisung des absoluten Wertes zu einem Register ausserhalb der Schleife würde $(N-1)-1=14$ Instruktionen einsparen, denn damit könnte Operation 8.2. parallel zu Zeile 7 ausgeführt werden, und damit Operation 9 parallel zu Operation 8.1. Offenbar führt der Compiler DSP-spezifische Schleifenoptimierungen, die sich aus der Parallisierbarkeit von Operationen ergeben, nicht durch.
- Die Modulo-Adressierung wird nicht durchgeführt. Stattdessen erfolgt in den Zeilen 8 und 9 das Umladen von Speicherbelegungen. Wie im vorherigen Punkt gesehen, könnte dieses Umladen parallel zu den arithmetischen Operationen durchgeführt werden, würde also im vorliegenden Fall keinen zusätzlichen Overhead erzeugen. Prinzipiell aber wäre der Compiler mittels einer aufwendigen Buchhaltung in der Lage zu erkennen, daß die in C als Kopieroperation ausgedrückte Funktion durch eine Modulo-Adressierung realisierbar ist: zu Beginn der Schleife ist bekannt, daß px und px2 um eins versetzt auf dasselbe Datenfeld zeigen. Innerhalb der Schleife werden die Dekrementierungen parallel fortgesetzt. Anhand der Kopieroperation kann auch festgestellt werden, daß ein Überschreiben der Werte, wie es die Modulo-Operation verlangt, erlaubt ist. Offensichtlich führt dieser Ansatz für den allgemeinen Fall zu einer hohen Komplexität des Compiliervorgangs. Weiterhin ist die Buchhaltung bei Verwendung der Zeiger-Adressierung nicht möglich, da man immer von Seiteneffekten ausgehen muß. Allerdings muß festgehalten werden, daß auch bei einer Feld-Adressierung die Möglichkeit der Modulo-Adressierung nicht genutzt wird. Keiner der in DSPstone untersuchten Compiler benutzt die Möglichkeit der Modulo-Adressierung.

Die Operation `y += *px * *ph` zerlegt der Compiler zunächst in eine Multiplikation (Zeile 6), eine anschliessende Linksverschiebung, so daß der Integer-Wert im hohen Register steht, sowie eine anschliessende Addition von 2 Integer-Zahlen. Durch dieses sequentielle Vorgehen ist sichergestellt, daß exakt die in C spezifizierte Integer-Arithmetik eingehalten wird. Alle durchgeführten Operationen basieren auf einer 16-bit Repräsentierung. Sobald sich ein Overflow ergibt, ist das Ergebnis falsch, der Benutzer hat bei der Programmierung für die korrekte Skalierung der Integer-Werte Sorge zu tragen.

Wird statt der sequentiellen Bearbeitung die `mac` Operation verwendet, so wird das Ergebnis mit 32-bit zwischengespeichert. Nach Abschluß der Akkumulation muß eine Verschiebung um 15 Stellen nach links erfolgen, um den Integer-Wert in das hohe Register abzubilden. Sobald also das Endergebnis nicht mehr durch 16 bit darstellbar ist, liegt auch hier ein falsches Ergebnis vor. Ergibt die Skalierung im sequentiellen Fall eine korrekte Berechnung, so liegt auch bei der `mac` Operation ein korrektes Ergebnis vor und beide Ergebnisse stimmen überein. Daher können die Zeilen 6-8 sowie 13-15 jeweils mittels einer `mac` Operation realisiert werden, ohne daß sich die Endergebnisse in der 16-bit-Genauigkeit unterscheiden.

Die bisher betrachteten Ineffizienzen sind also nicht prinzipieller Natur, denn die zur Optimierung notwendigen Informationen sind in der C-Formulierung bereits vorhanden und mit unterschiedlich hohem Aufwand extrahierbar. Der Compiler nutzt diese Informationen allerdings nicht aus.

Ein zusätzlicher Aspekt der Einschränkung wird hier nur indirekt deutlich. Da die NEC-Architektur auf einer Fractional-Darstellung der Festkomma-Werte basiert (d.h. der Wertebereich ist $[-1,1)$), ist in Assembly die effiziente Formulierung im Fractional-Bereich möglich. Da, wie oben angesprochen, eine Skalierung der Koeffizienten vorgenommen werden muß, kann diese die Zahlen auch direkt in den Fractional-Bereich überführen. Eine Multiplikation kann daher bei Verwendung der Fractional-Arithmetik nie zu einem Überlauf führen. Dieses vereinfacht die Anforderungen an die Skalierung wesentlich. Aus C heraus ist die Ausnutzung der Fractional-Darstellung nicht möglich, daher muß der Compiler immer wieder eine Verschiebung einfügen, um auf die Integer-Darstellung zurückzukehren. Das wiederum erfordert eine wesentlich restriktivere Skalierung. Für das FIR-Filter würde eine Programmierung mit fractional-Zahlen eine Einsparung von einer Instruktion im optimalen Assembly-Code bedeuten, der Compiler-generierte könnte auf 15 Instruktionen verzichten. Nicht erfaßt wird bei dieser Betrachtung der erzielbare Genauigkeitsgewinn durch die wegfallende Skalierung. Für eine genauere Diskussion der Auswirkung einer Fractional-Programmierung auf Genauigkeit und Skalierungsanforderungen sei auf [4] verwiesen.

Damit sind die beiden prinzipiellen Einflußgrößen für Compiler-Ineffizienzen bereits an einem einzigen Beispiel identifizierbar:

- nicht durchgeführte mögliche Optimierungen
  → der Informationsgehalt des Quell-Codes wird nicht ausgenutzt
- Restriktionen der Eingangssprache
  → der Wortschatz der Sprache ist unzureichend

3 Spracherweiterungen

Das Konzept der Spracherweiterungen basiert auf den beiden identifizierten Gründen für die Compiler-Ineffizienz und den daraus abgeleiteten folgenden Aussagen:

- C ist ungeeignet, spezifische DSP-Eigenschaften zu adressieren
- es ist für einen Compiler aufgrund der Komplexität der Aufgabe nicht möglich, den kompletten Informationsgehalt des Hochsprachenprogramms auszunutzen. Dieses liegt zum einen an der nicht akzeptierbaren Zunahme der Compilierzeit, zum anderen an der fehlenden Marktakzeptanz für hochoptimierende Compiler zu einem hohen Preis.

Soll dennoch nicht auf die Vorteile einer Hochsprachenprogrammierung verzichtet werden, lautet die Konsequenz:

Erlaube dem Programmierer die Mitteilung DSP-spezifischer Optimierungsmöglichkeiten, ohne dabei eine hardwarespezifische Programmierung vorzunehmen.

Dieser Anspruch führt auf zwei Ansätze, die im folgenden näher vorgestellt werden sollen.

3.1 Applikationsorientierte Erweiterungen

Diese Erweiterungen orientieren sich an den typischen Applikationen, die auf einem DSP ausgeführt werden sollen.

Eine typische Applikation auf dem DSP ist beispielweise die ausführlich behandelte FIR-Filterung. Eine applikationsbezogene Erweiterung muß eine Operation FIR zur Verfügung stellen, deren Operanden der Daten- und der Koeffizientenvektor sind. Es handelt sich also um einen funktional orientierten Ansatz.

Derzeit existieren zu jedem Prozessor umfangreiche Assembly-Bibliotheken, die genau diese Funktionalitäten enthalten, die hier als Spracherweiterungen vorgeschlagen werden. Wo liegt also der Vorteil der Erweiterungen? Messungen zeigen, daß die Verwendung von Assembly-Bibliotheken zu einem Laufzeit-Verlust von ca. 50% - 100% relativ zum optimalen Code führt. Dieses liegt insbesondere daran, daß die Assembly-Funktionalität generisch beschrieben werden muß, um für eine Vielzahl von Applikationen einsetzbar zu sein. Im Gegensatz dazu erfolgt die Compilierung der Erweiterungen kontextabhängig. Der Compiler kann also unterschiedlichen, an die Umgebung angepaßten Assembly-Code generieren. Weiterhin besteht nicht mehr die Notwendigkeit, zunächst Register und Pipeline zu sichern, bevor die Assembly-Funktionalität eingebunden wird.

Weitere Möglichkeiten bietet die Einführung neuer Datentypen, die sich nicht an der Hardware orientieren, sondern an der zu beschreibenden Aufgabenstellung. Hier sind z.B. die Einführung komplexer Zahlen sowie Vektoren zu erwähnen. Der Compiler kann aus der Kenntnis der Datentypen die auf diesen Datentypen auszuführenden Operationen effizienter umsetzen als nach der künstlichen Abbildung auf in C definierte Operationen.

Der entscheidende Vorteil dieses Ansatzes ergibt sich aus der gemeinsam erzielbaren Steigerung sowohl der Implementierungseffizienz als auch der Modellierungseffizienz. Dabei verstehen wir unter der Modellierungseffizienz die Möglichkeit, ein Problem effizient zu beschreiben. Die Implementierungseffizienz beschreibt die Qualität des Compiler-generierten Codes.

Es existieren bereits einige Konzepte, die unserer Definition der applikationsorientierten Erweiterungen entsprechen. Numerical C von Analog Devices [5] beispielsweise beinhaltet den Datentyp 'complex', sowie den Operator 'sum', der die Summation von Array-Elementen durchführt. DSPL [6] erlaubt die Mitteilung, daß eine FIR-Filterstruktur vorliegt.

Ein wesentliches Problem stellt die Definition der relevanten Funktionalitäten dar. Wenn eine neues Problem nicht effizient unter Ausnutzung der definierten Funktionalitäten formuliert werden

kann, bleibt nur die Programmierung in C. Es besteht also nicht die Möglichkeit einer Ausnutzung der spezifischen Architektur-Eigenschaften.

Das Konzept der applikationsorientierten Erweiterungen von C scheint gerechtfertigt, wenn man dem Umstand Rechnung tragen will, daß auch die DSP-Architektur applikationsbezogen aufgebaut ist. Bestes Beispiel ist die Unterstützung des Bit-Reversed-Adressierungs-Modus, der allein für die effiziente Implementierung der FFT eingefügt wurde. Es ist daher nicht einzusehen, warum diese Eigenschaft nicht auch auf die Programmierungsebene erweitert werden sollte.

3.2 Architekturspezifischer Ansatz

Der Ansatz der architekturspezifischen Spracherweiterungen löst sich von dem bisher verfolgten Ansatz, die Formulierung eines Problems unabhängig von der Zielarchitektur durchzuführen. Er basiert auf der Erfahrung, daß im allgemeinen die Architektur-Klasse (μP, DSP oder GPP) bereits zu Beginn eines Projekts feststeht. Daher soll der Entwickler die Möglichkeit erhalten, seine architekturspezifischen Kenntnisse in die Programmierung einzubringen, ohne damit eine prozessorspezifische Programmierung vorzunehmen.

Mit dieser Anforderung ist zunächst die Identifikation der allgemeinen Eigenschaften der DSP-Architekturen notwendig, die dann auf ein abstraktes Prozessormodell führen, das Grundlage der Programmierung der Klasse der DSPs werden soll.

Im folgenden soll exemplarisch der Ansatz sowie die Konsequenzen für die Hochsprache C diskutiert werden.

Die Speicherorganisation basiert auf einer Harvard- oder modifizierten Harvard-Architektur. Das erlaubt in der Regel nicht nur eine Trennung von Programm- und Datenworten, sondern auch das getrennte Ablegen von Datenworten in verschiedene Speicherbereiche. Damit sollten die Spracherweiterungen einen Hinweis auf die geeignete Ablage von Daten in bestimmten Speicherbereichen zulassen. Weiterhin kann für DSPs i.a. nicht von einer homogenen Speicherarchitektur ausgegangen werden, vielmehr wird ein Teil des Speichers auf dem Chip, ein anderer extern angeordnet sein, was sich zumindest in unterschiedlichen Zugriffszeiten ausdrücken kann. Somit sollte es dem Entwickler möglich sein, zu spezifizieren, ob Daten möglichst intern anzuordnen sind. Dieses Konzept der Speicherinformation stellt keine fundamentale Erweiterung von C dar, denn schon jetzt kann z.B. durch den 'storage class specifier' `register` innerhalb einer Deklaration dem Compiler mitgeteilt werden, daß es vorteilhaft ist, eine Variable möglichst lange im Register zu halten.

Weiterhin besitzen nahezu allen modernen DSP-Architekturen die Möglichkeit der Modulo- Adressierung. Der Benutzer sollte daher die Möglichkeit erhalten, explizit eine Modulo-Adressierung vorzunehmen. Dieses stellt ebenfalls keine prinzipielle Neuerung für die Sprache dar, denn die Inkrementierung/Dekrementierung von Zeigern ist ein zentrales Element von C. Nun ergibt sich als Unterschied, daß bei der Definition eines Feldes mitgeteilt werden muß, ob für dieses Feld eine zirkulare Adressierung vorzusehen ist. Allerdings ist die Anzahl der Ringspeicheradressierungen oft durch die Architektur begrenzt. Der Compiler erhält dann die Aufgabe, eine andere Speichermöglichkeit zu realisieren. I.a. ist es einfacher, aus der Information über die Möglichkeit einer Modulo-Adressierung eine alternative Darstellung zu finden, als beliebige alternative Darstellungen auf ihre Eignung zur Modulo-Adressierung zu untersuchen.

Bit-Reversed-Adressierung ist ein weiterer Adressierungsmodus, der von der Klasse der DSPs untersützt wird. Dieser Modus ist insbesondere durch die FFT motiviert, für die so eine effiziente Inplace-Verarbeitung möglich wird. Hier umfaßt die notwendige Information für den Compiler die Inkrementierung des Zeigers (vor der Bit-Umkehrung) sowie die Information über das Feld, auf das im Bit-Reversed-Mode zugegriffen werden soll.

Ein DSP besitzt keine allgemeine Wortbreite für Speicher, Bus und Register, sondern vielmehr eine heterogene Struktur. Es muß damit dem Benutzer möglich sein, zu jedem Zeitpunkt seine Anforderungen an die Darstellung einer Variablen anfügen zu können.

Eine detaillierte Entwicklung des abstrakten Prozessor-Modells wird in einer folgenden Veröffentlichung behandelt werden.

Es existieren bereits einige Ansätze der architekturspezifischen Erweiterungen. Im folgenden Abschnitt wenden wir uns dem NEC-Konzept [7] zu. Darüber hinaus erlaubt Numerical C die explizite Zuweisung einer Variablen zum Daten- oder Programmspeicherbereich. Das gleiche gilt für den Tartan-Compiler für die TI-Fließkomma-Familie, der eine direkte Unterstützung der Modulo-Adressierung erlaubt.

3.3 Vergleich der beiden Erweiterungs-Ansätze

Offensichtlich bieten beide Ansätze dem Compiler die Möglichkeit, DSP-spezifische Eigenschaften gezielter auszunutzen. Der mögliche Effizienzgewinn hängt also wesentlich davon ab, wie der Compiler die zusätzliche Information verarbeiten kann.

Grundsätzlich ist es für den Compiler einfacher, die architekturspezifischen Informationen zu berücksichtigen, da diese häufig eine direkte Umsetzung in Assembly ermöglichen. Dieser Ansatz ist sicherlich allgemeingültiger, da allen Applikationen die Architekturinformation zur Verfügung steht. Für den Programmierer bedeutet dieser Ansatz eines Herabsetzung der Abstraktion, denn schon zur Programmierzeit muß das, allerdings abstrakte, Prozessormodell berücksichtigt werden.

Der applikationsspezifische Ansatz erfordert hingegen vom Compiler eine höhere Komplexität. Der Programmierer hat die Möglichkeit, auf dem Abstraktionsniveau der Hochsprache weiterzuarbeiten und kann auf die Erlernung des Prozessormodells verzichten. Hier muß die Funktionalität der erweiterten Sprache erlernt werden.

Der wesentliche Vorteil des applikationsbezogenen Ansatzes liegt somit in der gleichzeitigen Verbesserung der Modellierungs- und Implementierungseffizienz der Programmierung. Der wesentliche Vorteil der architekturspezifischen Erweiterung liegt hingegen in der Allgemeinheit des Konzepts sowie in der geringeren notwendigen Compiler-Komplexität.

4 Ein Beispiel: Spracherweiterungen für den NEC7701x

Für den NEC7701x wird von Intermetrics das bisher umfangreichste Konzept der DSP-bezogenen Spracherweiterungen für die Sprache C angeboten. Es handelt sich dabei im wesentlichen um eine architekturbezogenes Erweiterungskonzept. Dieses soll hier nicht im Detail vorgestellt werden (dazu wird auf [7] verwiesen), vielmehr sollen die Handhabung und die sich ergebenden Konsequenzen für den Compiler-generierten Code anhand des FIR-Beispiels diskutiert werden.

Wie bereits bei der Diskussion in Abschnitt 2 identifiziert, ergaben sich bei der ANSI-C-Programmierung Ineffizienzen insbesondere aus der nicht optimalen Speicherzuweisung. Offensichtlich ist es vorteilhaft, Daten- und Koeffizientenvektoren in unterschiedlichen Speicherbänken abzulegen. Weiterhin ist es vorteilhaft, auf den Datenvektor mittels der Modulo-Adressierung zuzugreifen, da so ein Umspeichern vermieden werden kann. Die Typdeklaration als Integer wurde als Grund dafür identifiziert, daß die `mac` Operation nicht genutzt werden konnte.

Die C-Erweiterungen ermöglichen es nun, einer Variablen explizit mitzuteilen, in welcher Speicherbank sie abzulegen ist. Es handelt sich also im Unterschied zu den im vorherigen Abschnitt angeregten allgemeinen Spracherweiterungen um keine Empfehlung, sondern um eine feste Anweisung. Dieses geschieht durch eine Erweiterung des Typdeklarators um `_x, _y`. Natürlich wird bei der Verwendung der Erweiterungen vom Programmierer erwartet, daß er die Vorteilhaftigkeit der Zuordnung für bestimmte Operationen kennt.

Modulo-Adressierung wird ermöglicht durch die Bereitstellung eingebauter Funktionen, die vom Compiler direkt unterstützt werden. Es existieren 6 Funtionen, die auf den Datentyp `circ_buff` angewendet werden können:

`_Init_cb( cb, ptr, stride, bound)`
Initialisierung des Ringspeichers: Anfangsadresse (ptr), Inkrementierung/Dekrementierung (stride), Größe (bound)

`_Get_cb (cb, type)`
Ausgabe des Wertes, auf den der Ringspeicherzeiger gerade zeigt, auf den Datentypen `type` abgebildet.

`_Next_cb(cb, type)`
Wie `_Get_cb(cb, type)`, allerdings wird der Zeiger entprechend der Einstellung in `stride` weitergeschoben.

`_Tell_cb(cb), _Set_cb(cb, index)`
Ausgabe und Setzen der derzeitigen Zeiger-Stellung.

`_Set_stride(cb, stride)`
Setzen der Inkrementierung/Dekrementierung auf einen neuen Wert.

Offensichtlich muß sich der Benutzer vor der Anwendung zunächst mit den Funktionalitäten vertraut machen. Die Programmierung erfordert keine hardwarespezifischen Kenntnisse, sondern lediglich ein Verstehen des abstrakten Ringspeichermodells.

Neben diesen Hinweisen für den Compiler, die geeignet sind, die Komplexität herabzusetzen, existieren auch einige Erweiterungen, die über das in C ausdrückbare hinausgehen. So kann eine Variable als `accum` deklariert werden, d.h. sie soll mit der maximalen Wortbreite von 40-bit dargestellt werden. Weiterhin wird der Datentyp `fractional` eingeführt, der eine Verarbeitung von Zahlen im Bereich $[-1, 1)$ erlaubt und damit der implementierten Arithmetik entspricht.

Unter Ausnutzung dieser Erweiterungen ergibt sich folgender Code für das FIR-Filter:

```
y = 0;
data = _Next_cb (x_buff, fixed);
coeff = _Next_cb (c_buff, fixed);

for (i = 0; i < LENGTH - 2; i++)
  {
    y += data * coeff;
    data = _Next_cb (x_buff, fixed);
    coeff = _Next_cb (c_buff, fixed);
  }

y += data * coeff;
data = _Get_cb (x_buff, fixed);
coeff = _Next_cb (c_buff, fixed);

y += data * coeff;
_Get_cb (x_buff, fixed)= next;
```

Dabei wurde der Datentyp von y zu `accum` deklariert. Der Assembly-Code lautet:

```
1  CLR (r5)    r0=*dp6%%  r1=*dp2%%;
2  REP  14;
3    r5=r5 + r1h*r0h  r0=*dp6%%  r1=*dp2%%;
4  r5=r5 + r1h*r0h  r0=*dp6%%  r1=*dp2;
5  r5=r5 + r1h*r0h;
6  r0=*_next:X;
```

Dieser Assembly-Code ist äquivalent zum optimalen mit dem einizigen Unterschied, daß der nächste Wert direkt adressiert wird, was eine Parallelisierung der Zeilen 5 und 6 verhindert.

Es muß erwähnt werden, daß das erzielte Resultat stark von der C-Formulierung des Algorithmus abhängt. So erkennt man, daß im C-Code der erste Modulo-Zugriff außerhalb der Schleife erfolgt, analog zur Realisierung im Assembly-Code. Unsere erste Formulierung, die eine Schleife der Größe 15 vorsah mit dem Zugriff auf Daten und Koeffizienten vor der Akkumulation, resultierte in einer `loop` Operation mit 2 Instruktionen (was fast einer Verdopplung der Laufzeit entspricht), wobei in der ersten Instruktion die Daten geladen wurden, und in der zweiten Instruktion die `mac` Operation durchgeführt wurde. Offensichtlich findet keine Schleifen-Optimierung statt. Es muß vielmehr eine Programmierung in C erfolgen, die sich sehr stark an die endgültige Assembly-Version anlehnt.

Ein Vergleich der Laufzeiten und des Speicherbedarfs für die 3 Codeversionen des FIR-Filters ist in Fig.2 dargestellt.

Das hervorragende Ergebnis für die Realisierung des FIR-Filters kann natürlich nicht verallgemeinert werden. Innerhalb eines Projekts wurden von uns alle Kerne des DSPstone mit Spracherweiterungen realisiert. Dabei ergab sich in jedem Fall eine Verbesserung der erzielbaren Codequalitäten. Die Laufzeiten relativ zu denen der ANSI-C-programmierten lagen zwischen 19% (FIR-Filter) und 67% (n-real-update).

5 Spracherweiterungen: Assembly-Programmierung in der Hochsprache?

Wie aus den vorangegangenen Abschnitten zu erkennen ist, erlauben Spracherweiterungen eine wesentlich verbesserte Codegenerierung verglichen mit der Programmierung in ANSI-C. Bleiben aber die Vorteile einer Hochsprachenprogrammierung dabei nicht auf der Strecke? Anhand der eingangs

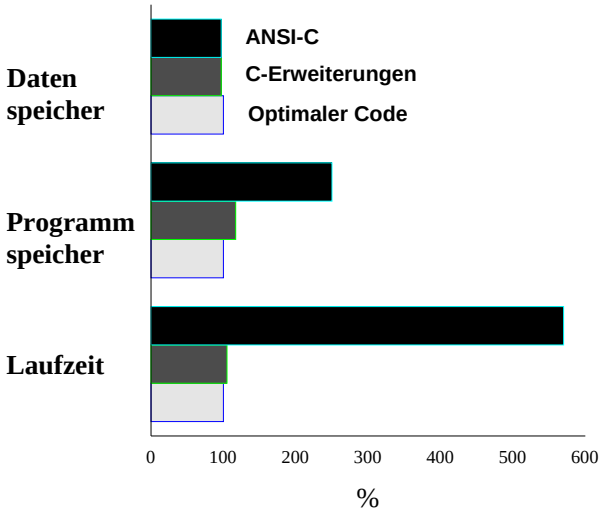

Abbildung 2: Relativer Laufzeit/Speicherbedarf für das FIR-Filter

vorgestellten Motivation soll untersucht werden, inwieweit sich durch die Einführung der Spracherweiterungen Veränderungen ergeben.

- Wartbarkeit:
  die Programmstruktur wird durch die Einführung von Spracherweiterungen nicht beeinflußt. Daher sind Änderungen und vor allem das Debugging mit den Vorteilen der Hochsprache weiterhin durchführbar.

- Wiederverwendbarkeit:
  Für den Code bestehen die gleichen Möglichkeiten hinsichtlich der Einbindung in Funktionen wie in der Hochsprache, auch hier ergibt sich keine Veränderung.

- Portierbarkeit:
  Offensichtlich bestehen hier wesentliche Probleme, solange kein Standard festgelegt ist. Wie unsere Darstellungen zum abstrakten Prozessormodell gezeigt haben, ist es möglich, auch architekturspezifische Erweiterungen zu definieren, die prinzipiell für alle Signalprozessor-Architekturen ausgenutzt werden können. Die derzeitigen Insellösungen sind nicht geeignet, auf breite Akzeptanz zu stoßen.

  Im vorherigen Abschnitt haben wir gesehen, daß für den dort vorgestellten Ansatz die Programmierung der spezifischen Informationen sehr nah am Assembly-Modell erfolgen muß, um den optimalen Code zu erzielen. Dieses entspricht unserer Erfahrung, daß unterschiedliche Compiler eine unterschiedliche C-Beschreibung zur Generierung von effizientem Code erfordern. Damit ist eine Portierbarkeit des Codes zwar gewährleistet, eine Portierung der Effizienz jedoch nicht.

- Produktivität:
  Eine Assembly-nahe Programmierung durch die Verwendung von Spracherweiterungen verringert die Programmiereffizienz wesentlich. Allerdings ist die Alternative, die reine Assembly-Programmierung verbunden mit dem inlining von Assembly-Abschnitten, mit einem wesentlich höheren Aufwand verbunden. Außerdem beeinflußt eine teilweise Assembly-Programmierung bisher angesprochenen Aspekte wesentlich.

  Die Assembly-nahe Programmierung ist darüberhinaus nicht zwingend notwendig. Wie am obigen Beispiel bereits dargestellt wurde, sollte der Compiler häufig mittels einfacher Optimierungen in der Lage sein, den Code wesentlich zu verbessern.

  Im Vergleich zur ANSI-C Programierung wird vom Programmierer zur Ausnutzung architekturspezifische Erweiterungen die Kenntnis des abstrakten Prozessormodells verlangt. Dieses stellt einen einmaligen Initialisierungsaufwand dar, der die Produktivität nur unwesentlich beeinflußt. Applikationsspezifische Erweiterungen können die Modellierungseffizienz wesentlich erhöhen. Daher ist nach der entsprechenden Lernphase mit einer Verbesserung der Produktivität zu rechnen.

6 Zusammenfassung und Ausblick

Im vorliegenden Artikel wurden die Motivation, die Möglichkeiten sowie die Einschränkungen für die Einführung von DSP-spezifischen Spracherweiterungen behandelt. Ausgehend von den mittels DSPstone ermittelten Ergebnissen wurde gezeigt, welche Ineffizienzen sich derzeit im Compiliervorgang ergeben. Eine Analyse der Gründe führte zu der Erkenntnis, daß viele Optimierungen prinzipiell während des Compiliervorgangs bereits erkennbar sind, wegen der damit verbundenden Komplexität aber vom Compiler nicht durchgeführt werden. Dieses beruht zum einen auf der damit stark ansteigenden Compilierzeit, zum anderen auf den fehlenden finanziellen Anstrengungen zur Entwicklung hochoptimierender Compiler für DSPs. Weiterhin erlaubt C als die untersuchte Hochsprache es nicht, die heterogene Struktur des DSP insbesondere durch die variierenden Wortbreiten zu unterstützen sowie hardware-spezifische Stärken auszunutzen.

Als ein Konzept zur Beseitigung dieser Ineffizienzen wurden DSP-spezifische Spracherweiterungen identifiziert. Ziel muß es sein, dem Entwickler die Möglichkeit zu geben, klassenspezifische Informationen in den Code einzubringen und so dem Compiler Hinweise zur effizienten Umsetzung der Funktionalität in die Maschinensprache zu geben. Dieses Konzept beruht auf der Idee, daß man einer Architektur, die für eine bestimmte Applikationsdomäne entwickelt wurde, auch eine spezielle Sprache, die auf diese Domäne zugeschnitten ist, zuweisen soll.

Das Beispiel der Erweiterung für den NEC7701x hat gezeigt, daß durch Spracherweiterungen signifikante Verbesserungen erzielt werden können. Die kompletten Ergebnisse unter Einbeziehung der Spracherweiterungen werden in Kürze in einem umfassenden Report herausgegeben. Dieser Report wird die Benchmarking-Ergebnisse aller relevanten derzeit verfügbaren Compiler sowohl für Festkomma- als auch für Fließkomma-Prozessoren enthalten.

Effiziente Compiler werden auf die Möglichkeit der Spracherweiterungen in absehbarer Zeit nicht verzichten können. Dieses erfordert jedoch die Bereitschaft, sich auf einen Sprachstandard zu verständigen. Insellösungen werden von den Anwendern nicht akzeptiert werden. Weiterhin ersetzen Spracherweiterungen nicht die Notwendigkeit guter Optimierungsstrategien, vielmehr sollten sie als Hilfsmittel verstanden werden, bestimmte Optimierungsrichtungen zu verfolgen. Erst das Zusammenspiel von Spracherweiterungen und angepaßten Optimierungsverfahren wird wesentliche Verbesserungen der Compiler-Ergebnisse bewirken, die dann zu einer breiteren Akzeptanz von Compilern im Anwenderbereich führen werden.

Literatur

[1] S. Harbison, "Uses and Misuses of C++ in DSP Application Development," in *Proc. of ICSPAT '94*, pp. 703 – 708, Oct. 1994.

[2] D. Crookes, D. Freeman, and C. Gostling, "Improving DSP Software Productivity Using an Optimizing Pascal Compiler for the AT & T DSP32C," *Microprocessing and Microprogramming*, pp. 239 – 242, 1989.

[3] V. Živojnović, J. Martínez, C. Schläger, and H. Meyr, "DSPstone: A DSP-oriented benchmarking methodology," in *Proc. of ICSPAT'94 - Dallas*, Oct. 1994.

[4] M. Willems, "Fractional vs. Integer Arithmetic - Comparing Scaling and Quantization Effects," *Internal Report, Aachen University of Technology*, 1995.

[5] M. Hoffman and A. Zatsman, "Numerical C," *Proceedings of DSPs in der Praxis, Munich 93*, pp. 161 – 174, Oct. 1993.

[6] A. Schwarte and H. Hanselmann, "The Programming Language DSPL," in *Proc. Int. Conference on Intelligent Motion, Munich, 1990*, pp. 268 – 280, Jun. 1990.

[7] B. Krepp, "DSP Extensions to ANSI C," in *Proc. of ICSPAT '94*, pp. 695 – 702, Oct. 1994.