Nat64 und Casting für IML

Marco Romanutti[1,2] und Benjamin Meyer[1,2]

[1] *Fachhochschule Nordwestschweiz FHNW, Brugg*
[2] Schlussbericht

Im Modul Compilerbau wird eine Erweiterung für die bestehende Sprache IML spezifiziert und implementiert. Die Implementierung beinhaltet einen neuen Datentyp für natürliche Zahlen, sowie eine Möglichkeit den Datentyp int64 in den neuen Datentypen zu casten und umgekehrt.

1 Erweiterung

1.1 Einleitung

Unter natürlichen Zahlen werden die positiven, ganzen Zahlen und 0 verstanden:

$$\mathbb{N} = \{0; 1; 2; 3; \ldots\}$$

Die IML soll um einen neuen Datentyp `nat64` erweitert werden. Der neue Datentyp soll solche positiven, ganzen Zahlen bis Länge 64 in Binärdarstellung abbilden können. Es sollen die bestehenden Operationen unterstützt werden. Ausserdem soll ein explizites Casting zwischen dem bestehenden Datentyp `int64` und dem neuen Datentyp `nat64` möglich sein.

1.2 Lexikalische Syntax

Für den neuen Datentyp wird das Keyword `(TYPE, NAT64)` und ein Castingoperator hinzugefügt.

```
Datentyp:   nat64    (TYPE, NAT64)
Brackets:   [ ]      LBRACKET, RBRACKET
```

Casting ist nur von `(TYPE, INT64)` zu `(TYPE, NAT64)` und umgekehrt möglich. Als Castingoperator werden rechteckige Klammern (nachfolgend Brackets gennant) verwendet. Innerhalb der Brackets befindet sich der Zieldatentyp [1].

[1] zum Beispiel `[int64]`

1.3 Grammatikalische Syntax

Das nachfolgende Code-Listing zeigt, wie der neue Datentyp `nat64` eingesetzt werden kann.

```
// Deklaration
var natIdent1 : nat64;
var natIdent2 : nat64;
var natIdent3 : nat64;

// Initialisierung
natIdent1 init := [nat64] 50;
natIdent2 init := [nat64] 10;
natIdent3 init := natIdent1 + natIdent2;

// Casting von int64 nach nat64
var intIdent1 : int64;
intIdent1 init := 30;
natIdent3 := [nat64] intIdent1;

call functionWithNatParam([nat64] intIdent1);

// Casting von nat64 nach int64
var intIdent2 : int64;
intIdent2 init := [int64] natIdent3;

call functionWithIntParam([int64] natIdent3);
```

Literale werden standardmässig als `int64` interpretiert - ein `nat64`-Literal bedingt vorab deshalb den Castingoperator. Falls zwei Datentypen nicht gecastet werden können, wird ein Kompilierungsfehler geworfen. Folgendes Code-Listing zeigt ein solches Beispiel mit dem bestehenden Datentyp `bool`:

```
// Deklaration
var boolIdent : bool;
boolIdent init := false;
var natIdent : nat64;
// Throws type checking error:
natIdent init := [nat64] boolIdent
```

Unsere Erweiterung unterstützt keine impliziten Castings. Weitere Code-Beispiele sind im Anhang zu finden.

1.4 Änderungen an der Grammatik

Zusätzlich zu den bestehende Operatoren wurde ein neuer `castOpr` erstellt, welcher anstelle des Nichtterminal-Symbol `factor` verwendet werden kann.

```
castOpr := LBRACKET TYPE RBRACKET
```

Das bestehende Nichtterminal-Symbol `factor` wird um diese neue Produktion ergänzt:

```
factor := LITERAL
| IDENT [INIT | exprList]
| castOpr factor
| monadicOpr factor
| LPAREN expr RPAREN
```

1.5 Kontext- und Typen-Einschränkungen

Der `TYPE` zwischen `LBRACKET` und `RBRACKET` muss vom Datentyp `int64` oder `nat64` sein. Ein Casting zum Typ `bool` oder vom Typ `bool` zu `int64` resp. `nat64` führt zu einem Kompilierungsfehler.

Tabelle 1 zeigt die unterstützen Typumwandlungen der verschiedenen Datentypen. Typumwandlungen, welche zu potentiellen Laufzeitfehler führen, sind mit mit * gekennzeichnet. Der Datentyp `int64` umfasst einen Wertebereich von -2147483648 bis 2147483647, wobei das Most Significant Bit (MSB) für das Vorzeichen verwendet wird. Weil der Datentyp `nat64` nur positive, ganze Zahlen und die Zahl 0 darstellt, wird kein Vorzeichenbit benötigt. Der Wertebereich verschiebt sich dadurch auf 0 bis 4294967295. Falls bei Typumwandlungen der Wert ausserhalb des Wertebereichs des Zieldatentyps liegt, führt dies zu einem Laufzeitfehler. Bei der Umwandlung von `nat64` nach `int64` kann ein solcher Laufzeitfehler beispielsweise auftreten, falls es sich um einen Wert > 2147483647 handelt. Falls negative Werte von `int64` nach `nat64` umgewandelt werden, resultiert ebenfalls ein Laufzeitfehler.

Table 1: *Casting zwischen Datentypen*

Quell- \ Zieldatentyp	int64	nat64	bool
int64	✓	✓*	✗
nat64	✓*	✓	✗
bool	✗	✗	✗

2 Aufbau Compiler

Der Compiler basiert auf der IML (V2) und ist in Java geschrieben.

2.1 Scanner

Literale werden standardmässig als `int64` interpretiert - ein `nat64`-Literal bedingt vorab deshalb den Castingoperator. Vom Scanner werden Literale als `long` in Java eingelesen. Dieser kann Werte von -9223372036854775808 bis 9223372036854775808 annehmen und deckt somit den gesamten Wertebereich der beiden Datentypen `int64` und `nat64` ab. Die Überprüfung, ob der Wert innerhalb des gültigen Wertebereichs des jeweiligen Datentyps liegt, erfolgt zum Zeitpunkt der Code-Generierung.

2.2 Parser

Der neu eingeführte `castOpr` und die neue Produktion `factor := castOpr factor` müssen im Abstrakten Syntax Tree (AST) abgebildet werden. Abbildung 1 zeigt die Umwandlung von der konkreten in die abstrakte Syntax.

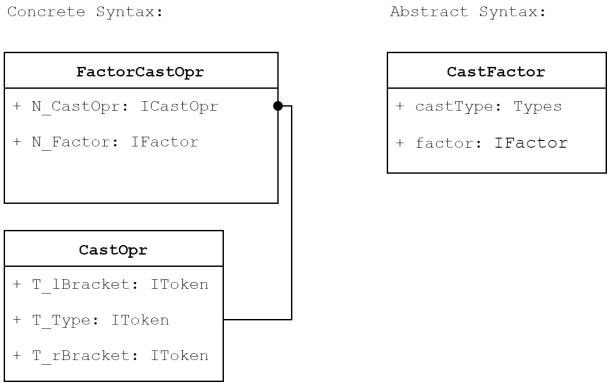

Figure 1: *Umwandlung von der konkreten in die abstrakte Syntax*

2.3 Statische Analyse

Scope checking

Für Routinen und Variablen liegen unterschiedliche Namespaces vor. Namen für Routinen und Variabeln können deshalb identisch sein. Es wird zwischen globalen und lokalen Namespaces unterschieden: Bei lokalen Namespaces werden die Variable innerhalb einer Routine definiert und haben dort ihre Gültigkeit. Globale Variabeln dürfen nicht denselben Namen haben wie lokalen Variabeln. Überladene Signaturen für Routinen[2] wurde in dieser Implementation nicht umgesetzt.

Bei *FunCallFactor, ProcCallCmd, DebugIn* und *AssignCmd* muss überprüft werden, ob die Parameter den richtigen `LValue`, resp. `RValue` besitzen. Folgende Kombinationen sind dabei allgemein erlaubt:

Bei einem *AssignCmd* muss der Ausdruck links zudem zwingend ein `LValue` sein. Bei einem *DebugIn* muss

[2] selber Name, unterschiedliche Parameterlisten

Table 2: *LRValue-Kombinationen*

Callee	Caller	Resultat
LValue	LValue	Valid
RValue	LValue	Valid (LValue dereferenzieren)
RValue	RValue	Valid
LValue	RValue	LRValueError ↯

es sich ebenfalls um einen `LValue` handeln, damit der Input-Wert dieser Variable zugewiesen werden kann.

Innerhalb des Scope checkings wird zudem überprüft, ob die Anzahl der erwarteten Parameter mit der Anzahl übergebener Parameter übereinstimmt.

Type checking

Das Casten zwischen zwei Datentypen ist nur für bestimmte Typen erlaubt (vgl. Tabelle 1). Zusätzlich sind bei der Abarbeitung des AST nur die folgenden Typen erlaubt:

Table 3: *Erlaubte Typen*

Klasse	Types
AddExpr, MultExpr, RelExpr	`int64, nat64` *
BoolExpr, IfCmd	`bool`
AssignCmd	`int64, nat64, bool` *
FunCallFactor, ProcCallFactor	Typ von Caller muss Typ von Callee entsprechen
MonadicFactor	NOTOPR: `bool` ADDOPR: `int64, nat64`
CastFactor	Typ von Factor und Typ von CastFactor müssen castable sein

Bei Einträgen, die mit * gekennzeichnet sind, müssen `LValue` und `RValue` vom selben Typ sein. Beim der Typenüberprüfung innerhalb vom `CastFactor` wird überprüft, ob der Typ vom `CastFactor` und jener des zugehörigen `factors` gecasted werden können (vgl. Tabelle 1). Die effektive Typ-Konversion wird erst bei der Code-Generierung durchgeführt.

Initalization checking

2.4 Virtuelle Maschine

Grundlage für die Code-Generierung ist der Abstract Syntax Tree (AST). Vom Root-Knoten ausgehend fügt jeder Knoten seinen Code zum Code-Array.

Im Falle eines Castings zwischen zwei Datentypen befindet sich an mindestens einer Stelle in der AST Struktur ein `CastFactor`-Element. Von diesem Element aus wird der Datentyp des zugehörigen `factor` geändert, indem dessen Attribut `castFactor` geändert wird. Dieses Attribut übersteuert den eigentlichen Datentyp des Elements innerhalb des AST. Weil der `factor` gemäss Grammatik unterschiedliche Produktionen besitzt (vgl. Änderungen an der Grammatik), muss die Typanpassung rekursiv weitergegeben werden. Die Rekursion wird durch Literale oder Expressions unterbrochen, wie am Beispiel in Anhang B aufgezeigt.

In der virtuellen Maschine wurde ein neuer generischer Typ `NumData` eingeführt. Dieser wird für die Konversion zwischen Daten vom Typ `IntData`[3] und `NatData`[4] verwendet. Abbildung 2 zeigt die Klassenhierarchie dieser Typen.

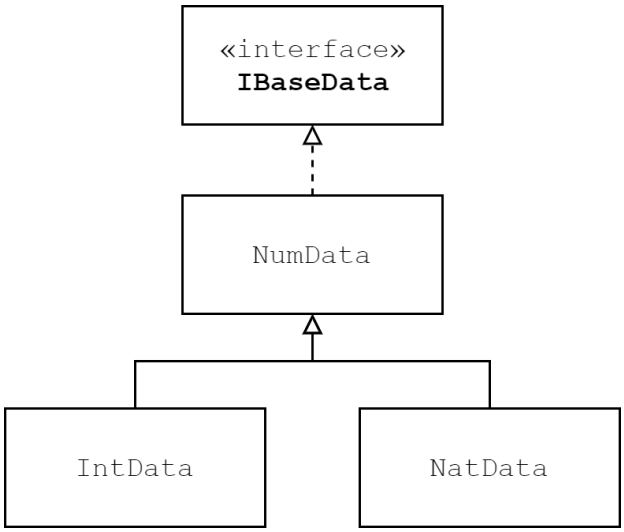

Figure 2: *Daten in VM*

Beim Dereferenzieren muss im Falle eines Castings der Datentyp von bereits typisierten Daten auf dem Stack geändert werden. Anhang C zeigt ein Beispiel, bei welchem der Datentyp aufgrund des neu propagierten Attributs `castType` erfolgreich umgewandelt wird.

2.5 Code Generierung

3 Vergleich mit anderen Programmiersprachen

3.1 Ganzzahlige Werte

In Java wird bei Zuweisungen die Länge einer Zahl in Bitdarstellung überprüft: Beim Datentyp `long` wird beispielsweise geprüft, ob der Wert als ganzzahliger Wert von 64-bit Länge dargestellt werden kann. Falls dies nicht der Fall ist, wird ein Fehler zur Kompilierungszeit geworfen. Das MSB wird als Vorzeichenbit verwendet, womit rund die Hälfte der vorzeichenlos darstellbaren `long`-Werte entfällt, resp. zur Darstellung von negativen Zahlen eingesetzt wird. Falls bei fortlaufenden Berechnungen Wertebereiche unter-

[3] für den Datentyp `int64`
[4] für den Datentyp `nat64`

resp. überschritten werden, führt dies zu einem arithmetischen Überlauf. Abbildung 3 zeigt den Überlauf bei ganzzahligen, vorzeichenbehafteten Datentypen (am Beispiel von Bitlänge 3 + 1).

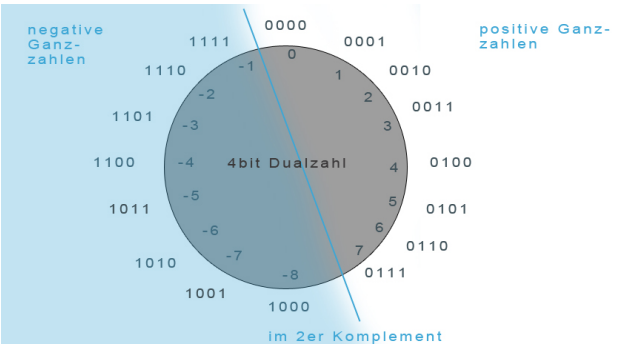

Figure 3: *Überlauf mit Integerzahlen*

Dadurch führt z.B. beim Datentyp `int` der Ausdruck `Integer.MAX_VALUE + 1` zum Wert `Integer.MIN_VALUE`. Dies kann dazu führen, dass mit „falschen“ Werten gerechnet wird, ohne dass der Entwickler dies bemerkt.

3.2 Fliesskommazahlen

Im Gegensatz zur Darstellung im Zweierkomplement, welche für Integer-Typen in Java verwendet werden, werden Fliesskommazahlen intern nach IEEE Standard dargestellt. Anders als bei der Zweierkomplement-Darstellung sieht dieses Format spezielle, konstante Werte vor. So sind in Java beispielsweise Konstanten für `Double.POSITIVE_INFINITY`, `Double.NEGATIVE_INFINITY` und `Double.NaN` definiert.

4 Designentscheidungen

4.1 Spezifiziertes Verhalten

Der neue Datentyp `nat64` unterstützt die bestehenden Operationen aus IML[5]. Sofern sich die einzelnen Operanden und auch das Resulat im Wertebereich[6] befinden, entspricht das Verhalten vom Datentyp `nat64` jenem vom Datentyp `int64`. Andernfalls wird folgendes Verhalten festgelegt:

- **Wertebereich**: Wertebereichsüber- resp. Unterschreitungen resultieren in einem Laufzeitfehler[7]. Dies erhöht die Typsicherheit beim Einsatz der verschiedenen Datentypen.
- **Nachkommastellen**: Gemäss IML-Spezifikation sind als Literale nur ganzzahlige Werte erlaubt. Falls z.B. bei einer Division ein Rest resultiert, wird ein ganzzahliges Resultat zurückgegeben. Der Wert des Resultats ist abhängig von der gewählten Operation (`DivFloor, DivTrunc, etc.`).

4.2 Alternative Ansätze

Folgende weiteren Ansätze wurden für die Umsetzung der Erweiterung in Betracht gezogen:

- **Arithmetischer Überlauf:** Wertebereisüber- resp. unterschreitungen resultieren in einem arithmetischen Überlauf. Gegenüber der Darstellung in Abbildung 3 müsste kein negativer Wertebereich verwendet werden und das Addieren von $+1$ zum grössten Darstellbaren Wert des Datentyps `nat64` führt zum Wert 0. Weil auf ein Vorzeichenbit verzichtet werden kann, verdoppelt sich der Wertebereich gegenüber dem Datentyp `int64`. Nachteilig ist dabei, dass der Entwickler verantwortlich ist für das Einhalten der Wertebereichsgrenzen.
- **Vordefinierte Konstanten:** Ähnlich wie bei Fliesskommazahlen (vgl. Kapitel 3.2) könnten konstante Werte für z.B. `POSITIVE_INFINITY` und `NEGATIVE_INFINITY` vorgesehen werden. Das Verhalten bei Wertebereichsüberschreitungen müsste definiert werden. Nachteilig bei dieser Variante ist, dass der Wertebereich um die Anzahl solcher konstante Werte verringert wird.
- **Absolute Werte:** Bei dieser Variante wird bei negativen Werten deren Absolutwert verwendet. Von dieser initial angedachten Variante wurde abgesehen, weil das Verhalten schnell zu ungewollten Resultaten führen kann.

5 Beispielprogramme

Operation:

```
program progAddition
global
var x:nat64;
var y:nat64;
var r:nat64;
var b:bool
do
x init := 4;
y init := 3;
r init := x + y;
b init := r = 7;

debugout r;
debugout b
endprogram
```

[5] Aktuell sind dies

- MULTOPR(*, divE, modE)
- ADDOPR(+, -)
- RELOPR(<, <=, >, >=, =, /=)
- BOOLOPR(, ||, , |)

[6] $[0, 4294967295]$

[7] Negative Zahlen entsprechen Wertebereichsunterschreitungen

Casting:

```
program progCasting
global
var x:nat64;
var y:int64;
var r:nat64;
var b:bool
do
x init := 4;
y init := 3;
r init := x + [nat64] y;
b init := r = 7;

debugout r;
debugout b
endprogram
```

References

[1] Wikipedia: Natürliche Zahl, https://de.wikipedia.org/wiki/Nat\%C3\%BCrliche_Zahl

[2] Wikipedia: Natural numbers (engl.), https://en.wikipedia.org/wiki/Natural_number

A Vollständige Grammatik

B Beispiel doppeltes Casting

IML:

```
program progDouble
global
var value:int64
do
value init := [int64] [nat64] ((4 + 1) + 1)
endprogram
```

Code-Array:

```
0: AllocBlock(1)
1: UncondJump(2)
2: LoadAddrAbs(0)
3: LoadImInt(4)
4: LoadImInt(1)
5: AddInt
6: LoadImInt(1)
7: AddInt
8: Store
9: Stop
```

UML:

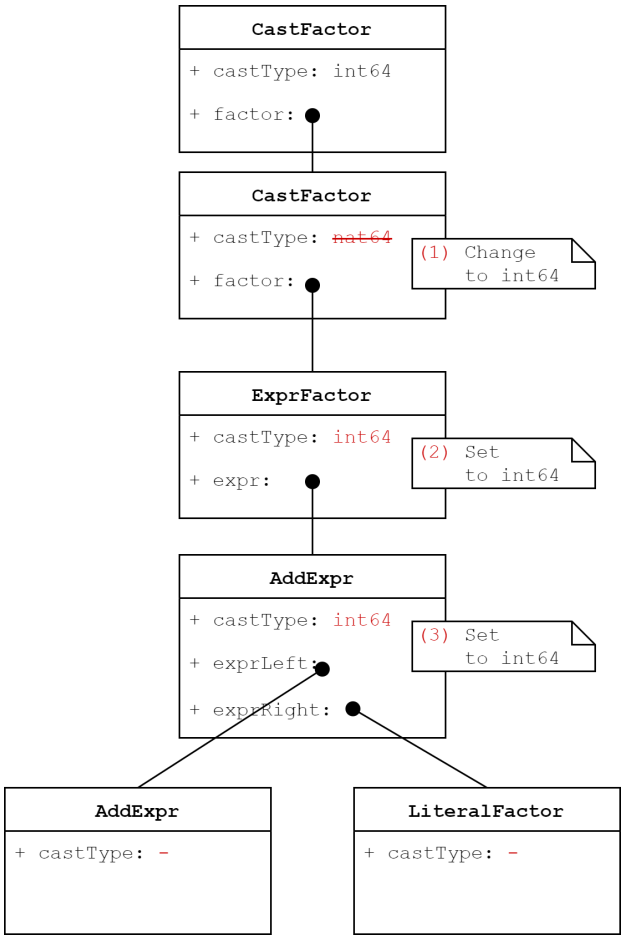

Figure 4: *Auszug aus AST*

C Beispiel `Deref`

IML:

```
program exampleCasting
global
var x:int64;
var y:nat64
do
x init := [int64] 4;
y init := [nat64] 3;
x := [int64] y
endprogram
```

Code-Array:

```
codeArrayPoiner: 14
0: AllocBlock(1)
1: AllocBlock(1)
2: UncondJump(3)
3: LoadAddrAbs(0)
4: LoadImInt(4)
5: Store
6: LoadAddrAbs(1)
7: LoadImNat(3)
8: Store
9: LoadAddrAbs(0)
10: LoadAddrAbs(1)
11: Deref
12: Store
13: Stop
```

Stack:

```
...

(8) Executing instruction Store
pc: 8
sp: 4
Stack content:
0: IntData: 4
1: NatData: 3

(9) Executing instruction LoadAddrAbs(0)
pc: 9
sp: 2
Stack content:
0: IntData: 4
1: NatData: 3
2: IntData: 0

(10) Executing instruction LoadAddrAbs(1)
pc: 10
sp: 3
Stack content:
0: IntData: 4
1: NatData: 3
2: IntData: 0
3: IntData: 1

(11) Executing instruction Deref
pc: 11
sp: 4
Stack content:
0: IntData: 4
1: NatData: 3
2: IntData: 0
3: IntData: 3

(12) Executing instruction Store
pc: 12
sp: 4
Stack content:
0: IntData: 3
1: NatData: 3

...
```