

Nat64 und Casting für IML

Marco Romanutti^{1,2} und Benjamin Meyer^{1,2}

¹Fachhochschule Nordwestschweiz FHNW, Brugg

²Zwischenbericht

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 Aufbau Compiler

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

1.1 Statische Analyse

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

Table 1: *LRValue-Kombinationen*

Callee	Caller	Resultat
LValue	LValue	Valid
RValue	LValue	Valid (LValue dereferenzieren)
RValue	RValue	Valid
LValue	RValue	LRValueError \neq

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

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

2 Erweiterung

2.1 Einleitung

Unter natürlichen Zahlen werden die positiven, ganzen Zahlen und 0 verstanden. Die IML soll um einen neuen Datentyp `nat64` erweitert werden. Der neue Datentyp soll solche positiven, ganzen Zahlen mit 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.

2.2 Lexikalische Syntax

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

Datentyp:	<code>nat64</code>	<code>(TYPE, NAT64)</code>
Brackets:	<code>[]</code>	<code>LBRACKET, RBRACKET</code>

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

2.3 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
```

¹zum Beispiel `[int64]`

```

natIdent1 init := 50;
natIdent2 init := 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);

```

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 checker error:
natIdent init := [nat64] boolIdent

```

Unsere Erweiterung unterstützt keine impliziten Castings. Weitere Code-Beispiele sind in Kapitel 5 zu finden.

2.4 Änderungen an der Grammatik

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

```
castOpr := LBRACKET ATOMTYPE RBRACKET
```

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

```

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

```

2.5 Kontext- und Typen-Einschränkungen

Der ATOMTYPE 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 2 zeigt die unterstützten Typumwandlungen der verschiedenen Datentypen. Typumwandlungen, welche zu potentielltem Informationsverlust führen, sind mit * gekennzeichnet. Bei der Umwandlung von `nat64` nach `int64` kann ein Informationsverlust resultieren, weil beim Datentyp `int64` das Most Significant Bit (MSB) für das Vorzeichen verwendet wird (vgl. Kapitel 4). Falls Werte von `int64` nach `nat64` umgewandelt werden, geht die Information zum Vorzeichen verloren und der Wert wird als absoluter Wert interpretiert.

Table 2: Casting zwischen Datentypen

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

3 Vergleich mit anderen Programmiersprachen (am Beispiel von Java)

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

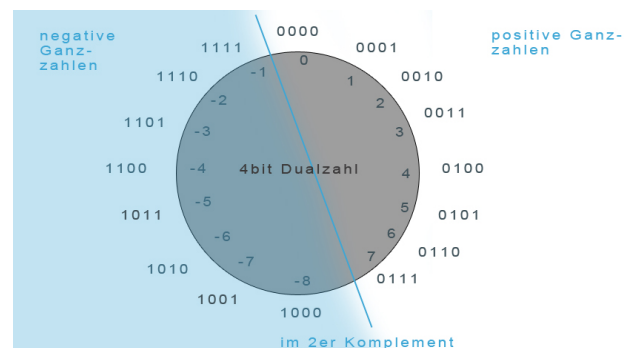


Figure 1: Überlauf mit Integerzahlen

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 Werte für `POSITIVE_INFINITY` und `NEGATIVE_INFINITY` vor.

4 Designentscheidungen

4.1 Spezifiziertes Verhalten

Der neue Datentyp `nat64` unterstützt die bestehenden Operationen aus IML². Sofern sich die einzelnen Operanden und auch das Resultat im Wertebereich ($\in \mathbb{N}$) befinden, entspricht das Verhalten vom Datentyp `nat64` jenem vom Datentyp `int64`. Andernfalls wird folgendes Verhalten festgelegt:

- **Wertebereich:** Bei einem Überlauf wird jeweils mit dem maximalen Wert weitergerechnet. Dieser entspricht dem maximalen Wert von `int64`³.
- **Negative Werte:** Werte werden jeweils als absolute Werte Betrachtet. Ein negativer Wert -5 entspricht beispielsweise dem Betrag, also $|-5| = 5$.
- **Rest bei Division:** Wird analog `int64` behandelt und Nachkommastellen werden abgeschnitten.

4.2 Alternative Ansätze

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

```

²Aktuell sind dies

- `MULTOPR(*, divE, modE)`
- `ADDOPR(+, -)`
- `RELOPR(<, <=, >, >=, =, /=)`
- `BOOLOPR(/&? ∨?)`

³9,223,372,036,854,775,807

`endprogram`

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