Floating Point Expansion in Coq

L. Théry

INRIA Sophia Antipolis

16 Novembre 2000

1 Motivation

What are expansions ?

Representing a multiple precision number:

$$78090027829990120$$

as list of floating-point numbers:

$$7.809 * 10^{16} + 2.783 * 10^{10} + -9.880 * 10^{4}$$

Usual operations (+,*,/) can be defined on expansions

Floating-point units are highly optimized

Base for adaptative arithmetic

Motivation

Developed a machine-checked version

Credits:

Marc Daumas

Claire Moreau-Finot

Sebastien Bardin

$\boxed{\text{Coq}}$

Prover based on type theory (Aldor):

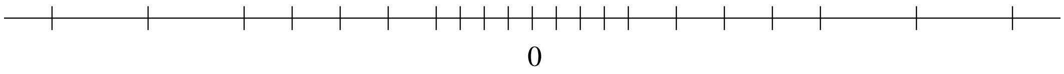
- Define new objects
- Define functions and predicates
- Prove properties (interactive construction)

Existing theories: $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$

+ Expressivity
+ Soundness
- Automation

4 Formalisation

Floating point numbers



$$m * r^e$$

Explicit representation of floating point numbers:

$$\{num : \mathbb{Z}, exp : \mathbb{Z}\} : \mathit{float}$$

$$n : \mathit{float} \to \mathbb{Z}$$

$$e : \mathit{float} \to \mathbb{Z}$$

$$p = q \iff n(p) = n(q) \wedge e(p) = e(q)$$

5 Formalisation

Floating point numbers

Semantic:

Parameter $r : \mathbb{N}$

Hypothesis: $1 < r$

Definition:

$$f2r : float \to \mathbb{R}$$

$$f2r(p) = n(p) * r^{e(p)}$$

Equality over $\mathbb{R}$ (implicit coercion)

$p == q$ interpreted as $f2r(p) == f2r(q)$

Properties:

$$p == 0 \iff n(x) = 0$$

$$p == q \wedge e(p) = e(q) \Rightarrow p = q$$

$$p == q \wedge n(p) = n(q) \Rightarrow p = q$$

6 Formalisation

$\boxed{\text{Shift}}$

Definition:

$$shift : \mathbb{N} \rightarrow float \rightarrow float$$

$$shift(n, p) = \{n(p) * r^n, e(p) - n)\}$$

Properties:

$$shift(n, p) == p$$

$$p == q \iff \exists n, m\colon \mathbb{N}.\, shift(n, p) = shift(m, q)$$

7 Formalisation

Comparison

$p < q$ interpreted as $f2r(p) < f2r(q)$

$p \leq q$ interpreted as $f2r(p) \leq f2r(q)$

Properties:

$$0 < n(x) \iff 0 < x$$

$$0 \leq n(x) \iff 0 \leq x$$

If $e(x) = e(y)$

$$x < y \iff n(x) < n(y)$$

$$x \leq y \iff n(x) \leq n(y)$$

Operations

Definition:

$$-p = \{-n(p),\, e(p)\}$$

$$|p| = \{|n(p)|,\, e(p)\}$$

If $m = min(e(p), e(q))$

$$p + q = \{n(p) * r^{e(p)-m} + n(q) * r^{e(q)-m},\, m\}$$

$$p - q = \{n(p) * r^{e(p)-m} - n(q) * r^{e(q)-m},\, m\}$$

$$p * q = \{n(p) * n(q),\, e(p) + e(q)\}$$

Properties:

$$f2r(-p) == -f2r(p)$$

$$f2r(|p|) == |f2r(p)|$$

$$f2r(p + q) == f2r(p) + f2r(q)$$

$$f2r(p - q) == f2r(p) - f2r(q)$$

$$f2r(p * q) == f2r(p) * f2r(q)$$

Bounded

Bound:

$$\{vNum : \mathbb{N}, dExp : \mathbb{N}\} : bound$$

$$v : bound \to \mathbb{N}$$

$$d : bound \to \mathbb{N}$$

To be bounded:

$$\mathcal{B}_b(p) \iff -v(b) \leq n(p) \leq v(b) \wedge -d(b) \leq e(p)$$

Properties:

$$\mathcal{B}_b(p) \iff \mathcal{B}_b(-p)$$

$$\mathcal{B}_b(p) \iff \mathcal{B}_b(|p|)$$

$$0 < p \wedge \mathcal{B}_b(p) \Rightarrow \{1, -d(b)\} \leq p$$

Sterbenz

Theorem:

$$\mathcal{B}_b(p) \wedge \mathcal{B}_b(q) \wedge q/2 \leq p \leq 2 * q \Rightarrow \mathcal{B}_b(p - q)$$

Proof:

Case $q \leq p \leq 2 * q$

Let $m = min(e(p), e(q))$

Let $p' = shift(e(p) - m,\ p)$

Let $q' = shift(e(q) - m,\ q)$

$p - q = \{n(p') - n(q'),\ e(p')\}$ and $e(p') = e(q')$

We have $q' \leq p' \leq 2 * q'$

$n(q') \leq n(p') \leq 2 * n(q')$

$0 \leq n(p') - n(q') \leq n(q')$

So $-v(b) \leq n(p - q) \leq v(b)$

$e(p - q) = e(p) \vee e(p - q) = e(q)$

So $-d(b) \leq e(p - q)$

Digits

On natural numbers: $digit : \mathbb{N} \rightarrow \mathbb{N}$

Properties:

$$digit_r(0) = 0$$

$$r^{q-1} \leq p < r^q \Rightarrow digit_r(p) = q$$

$$p \leq q \Rightarrow digit_r(p) \leq digit_r(q)$$

$$digit_r(p) < digit_r(q) \Rightarrow p < q$$

If $\neg(p == 0)$

$$digit_r(|n(shift(n, p))|) = digit_r(|n(p)|) + n$$

$$p == q \wedge digit_r(|n(p)|) = digit_r(|n(q)|) \Rightarrow p = q$$

$\boxed{\text{Precision}}$

Precision: $p_r : \mathbb{N}$

Hypothesis: $0 < p_r$ and $v(b) = r^{p_r} - 1$

Property:

$$digit(|n(p)|) \leq p_r \iff -v(b) \leq n(p) \leq v(b)$$

$\boxed{\text{Normal}}$

Definition:

$$\mathcal{N}(p) \Longleftrightarrow \mathcal{B}(p) \wedge digit(|n(p)|) = p_r$$

Properties:

$$\mathcal{N}(p) \Rightarrow \neg p == 0$$

$$\mathcal{N}(p) \iff \mathcal{N}(-p)$$

$$\mathcal{N}(p) \iff \mathcal{N}(|p|)$$

$$\mathcal{N}(p) \wedge \mathcal{N}(q) \wedge p == q \Rightarrow p = q$$

$$\mathcal{N}(p) \wedge \mathcal{N}(q) \wedge 0 \leq p < q \Rightarrow$$
$$e(p) < e(q) \vee (e(p) = e(q) \wedge n(p) < n(q))$$

Subnormal

Definition:

$$\mathcal{S}(p) \iff \mathcal{B}(p) \wedge digit(|n(p)|) < p_r \wedge e(p) = -d(b)$$

Properties:

$$\mathcal{S}(p) \iff \mathcal{S}(-p)$$

$$\mathcal{S}(p) \iff \mathcal{S}(|p|)$$

$$\neg(\mathcal{S}(p) \wedge \mathcal{S}(p))$$

$$\mathcal{S}(p) \wedge \mathcal{S}(q) \wedge p == q \Rightarrow p = q$$

$$\mathcal{S}(p) \wedge \mathcal{S}(q) \wedge p < q \Rightarrow n(p) < n(q)$$

Canonical

Definition:

$$\mathcal{C}(p) \Longleftrightarrow \mathcal{S}(p) \vee \mathcal{N}(p)$$

Properties:

$$\mathcal{C}(p) \iff \mathcal{C}(-p)$$

$$\mathcal{C}(p) \iff \mathcal{C}(|p|)$$

$$\mathcal{C}(p) \wedge \mathcal{C}(q) \wedge p == q \Rightarrow p = q$$

$$\mathcal{C}(p) \wedge \mathcal{C}(q) \wedge 0 \leq p < q \Rightarrow$$
$$e(p) < e(q) \vee (e(p) = e(q) \wedge n(p) < n(q))$$

Normalisation

Definition:

$$
\begin{array}{l}
norm(p) = \\
\quad \texttt{if } p == 0 \\
\quad \texttt{then } \{0,\ -d(b)\} \\
\quad \texttt{else } shift(min(p_r - digit(|n(p)|), d(b) + e(p)), p)
\end{array}
$$

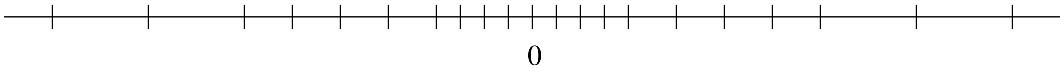
Propriétés:

$$\mathcal{N}(p) == p$$

$$\mathcal{B}(p) \Rightarrow \mathcal{C}(norm(p))$$

17 Formalisation

Successor



Definition:

$$
\begin{array}{l}
suc(p) = \\
\quad \texttt{if } n(x) = v(b) \\
\quad \texttt{then } \{r^{p_r - 1},\, e(p) + 1\} \\
\quad \texttt{else if } n(x) = -r^{p_r - 1} \\
\qquad \texttt{then if } e(p) = -d(b) \\
\qquad\quad \texttt{then } \{n(p) + 1,\, e(p)\} \\
\qquad\quad \texttt{else } \{-v(b),\, e(p) - 1\} \\
\qquad \texttt{else } \{n(p) + 1,\, e(p)\}
\end{array}
$$

Properties:

$$\mathcal{C}(p) \Rightarrow \mathcal{C}(suc(p))$$

$$\mathcal{C}(p) \Rightarrow p < suc(p)$$

$$\mathcal{C}(p) \wedge \mathcal{C}(q) \wedge p < q \Rightarrow suc(p) \leq q$$

Predecessor

Definition:

$$
\begin{array}{l}
pred(p) = \\
\quad \texttt{if } n(x) = -v(b) \\
\quad \texttt{then } \{-r^{p_r - 1},\, e(p) + 1)\} \\
\quad \texttt{else if } n(x) = r^{p_r - 1} \\
\quad\quad \texttt{then if } e(p) = -d(b) \\
\quad\quad\quad \texttt{then } \{n(p) - 1,\, e(p)\} \\
\quad\quad\quad \texttt{else } \{v(b),\, e(p) - 1\} \\
\quad\quad \texttt{else } \{n(p) - 1,\, e(p)
\end{array}
$$

Properties:

$$\mathcal{C}(p) \Rightarrow \mathcal{C}(pred(p))$$

$$\mathcal{C}(p) \Rightarrow pred(p) < p$$

$$\mathcal{C}(p) \wedge \mathcal{C}(q) \wedge p < q \Rightarrow p \leq pred(q)$$

Min and Max

Definition:

$$isMin(r, min) \iff$$
$$\mathcal{B}(min) \wedge \forall f\colon float.\, \mathcal{B}(f) \wedge f < r \Rightarrow f \leq min$$

$$isMax(r, max) \iff$$
$$\mathcal{B}(max) \wedge \forall f\colon float.\, \mathcal{B}(f) \wedge r < f \Rightarrow max \leq f$$

Properties

$$\forall r\colon \mathbb{R}.\, \exists f\colon float.\, isMin(r, f)$$

$$\forall r\colon \mathbb{R}.\, \exists f\colon float.\, isMax(r, f)$$

$$\forall r\colon \mathbb{R}.\, \forall f\colon float.\, isMin(r, f) \wedge \neg r == f \Rightarrow$$
$$isMax(r, suc(norm(f)))$$

$$\forall r\colon \mathbb{R}.\, \forall f\colon float.\, isMax(r, f) \wedge \neg r == f \Rightarrow$$
$$isMin(r, pred(norm(f)))$$

Rounding Mode

Abstract rounding: $P(p,\ q)$

q is a rounding value of p according to P

Definition:

$$RoundedMode(P) \iff$$
$$Total(P) \wedge Compatible(P) \wedge$$
$$MinOrMax(P) \wedge Monotone(P)$$

Properties:

$$\forall P. \forall f\colon float.\ RoundedMode(P) \Rightarrow P(p, p)$$

$$\forall p\colon \mathbb{R}.\ \forall q, q'\colon float.$$
$$\mathcal{B}(q') \wedge P(p, q) \wedge r \leq r * q' \Rightarrow q \leq r * q'$$

$$\forall x, z\colon float.\ \forall q, q'\colon float.$$
$$\mathcal{B}(x) \wedge Closest(y, z) \wedge x/2 < y \Rightarrow x/2 \leq z$$

$$\forall p, q, pq\colon float.\ \mathcal{B}(p) \wedge \mathcal{B}(q) \wedge Closest(p + q, pq)$$
$$\exists err\colon float.\ \mathcal{B}(err) \wedge err = (p + q) - pq$$

Fast Two Sum

$$
\begin{aligned}
&\mathit{FastTwoSum}(p, q) = \\
&\quad v_0 := p \oplus q \\
&\quad v_1 := v_0 \ominus p \\
&\quad v_2 := q \ominus v_1 \\
&\quad \mathtt{return}(v_0, v_2)
\end{aligned}
$$

Theorem:

$$
\begin{aligned}
&\forall p, q \colon \mathit{float}.\, \mathcal{B}(p) \wedge \mathcal{B}(q) \wedge e(q) \leq e(p) \Rightarrow \\
&\quad v_2 == p + q - v_0
\end{aligned}
$$

Development

Available at

`http://www-sop.inria.fr/lemme/AOC/coq`

84 Definitions 610 Theorems 11000 Lines

Conclusions

A relatively heavy infrastructure

General theorems

Proofs sensitive to our initial choice

Readable theorems

Unreadable proofs

$\boxed{\text{end}}$