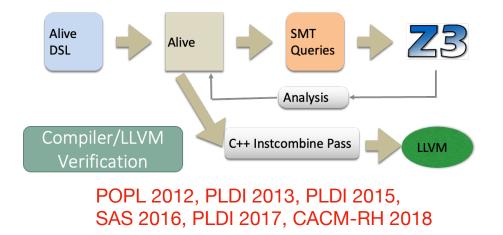
# A Case for Correctly Rounded Elementary Functions

Santosh Nagarakatte @ NJPLS 2022 Rutgers University

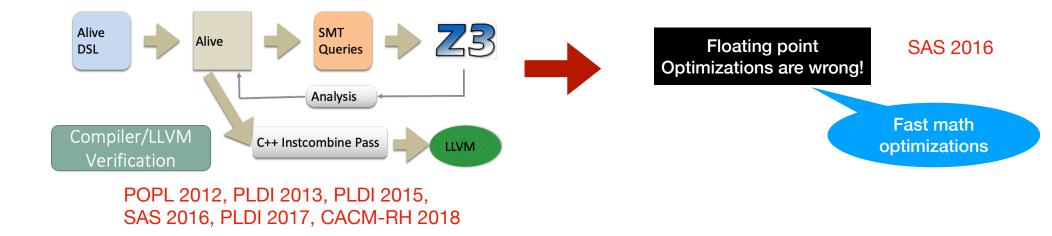
Collaborators: Jay Lim, Mridul Aanjaneya, John Gustafson, and Sehyeok Park

1

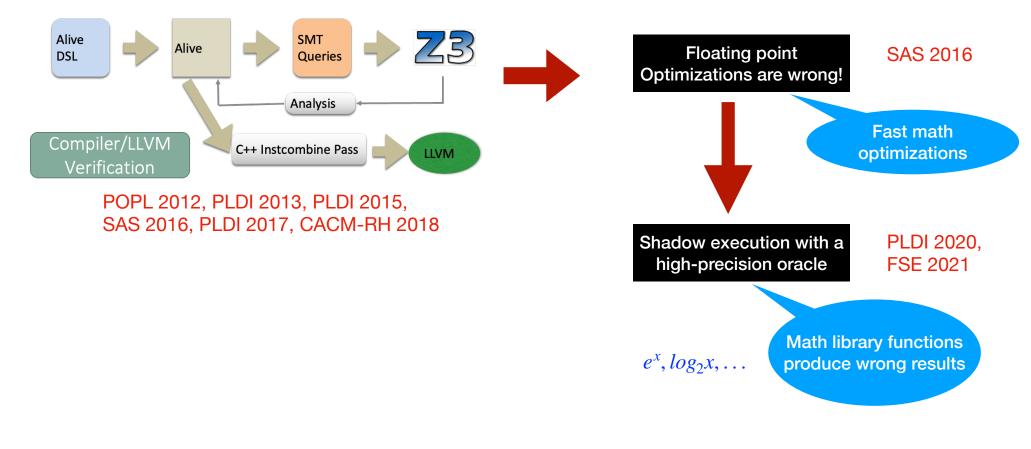






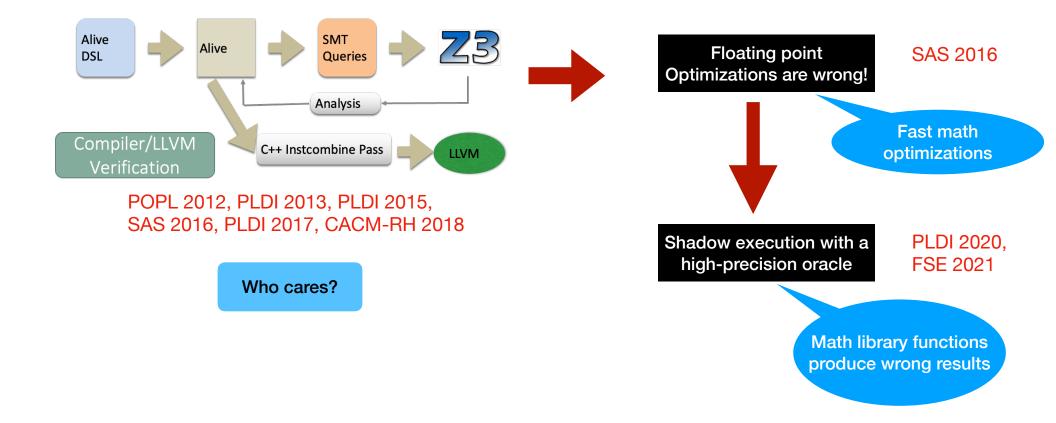


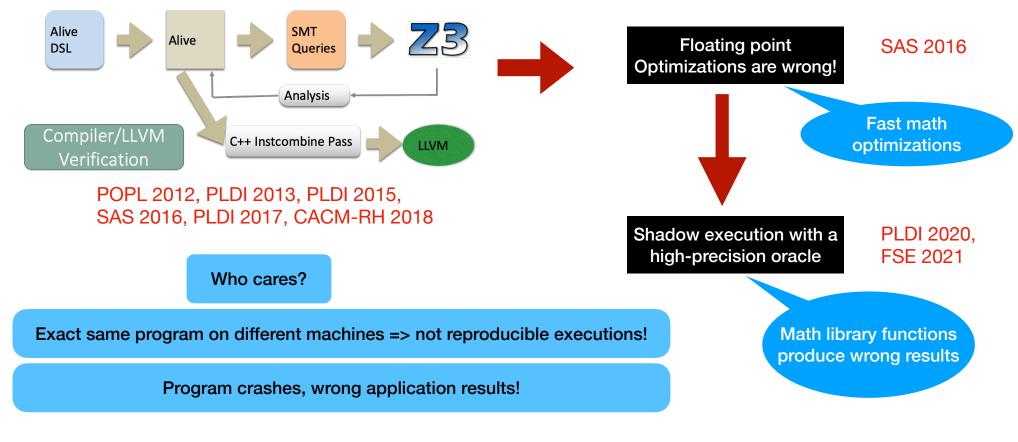




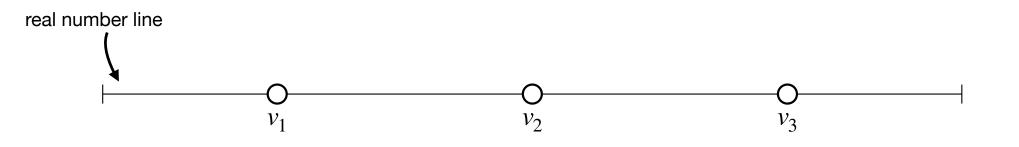
4

**RUTGERS** 

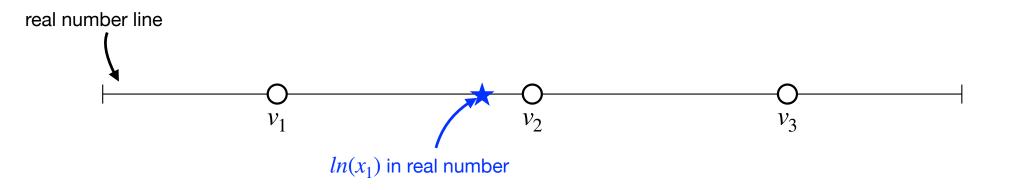




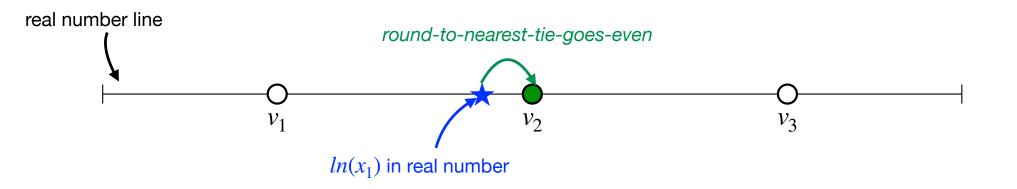
**RUTGERS** 



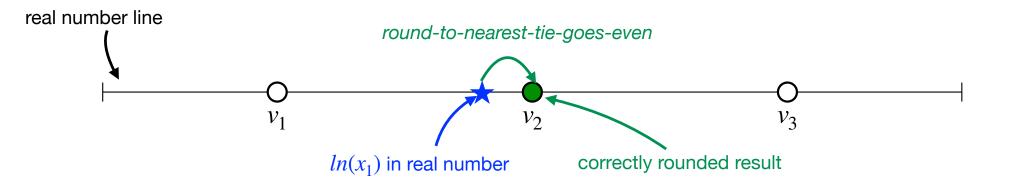






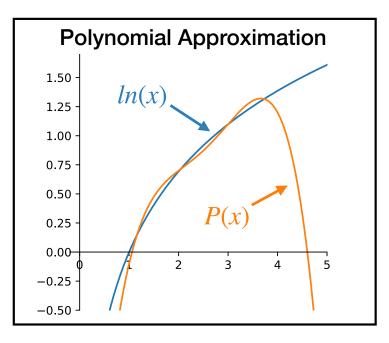








#### **How do Prior Techniques Approximate Elementary Functions?**



- 1. Approximate the REAL value of ln(x)
- 2. Feasible with small domains:

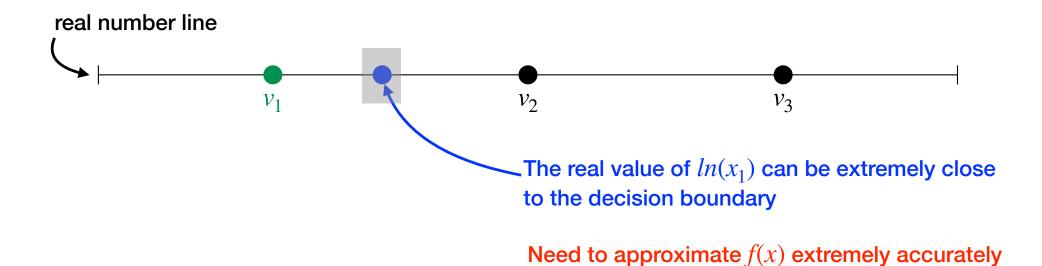
Range reduction to transform the input to a small domain

#### 3. Mini-Max Approximation:

Polynomial Approximation that minimizes the maximum error for all points

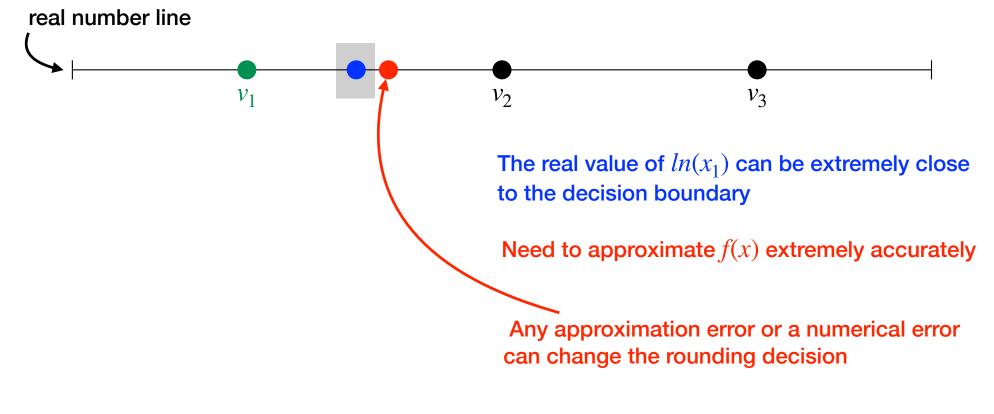
#### **RUTGERS**

#### What's the issue with Mini-Max Approximations?



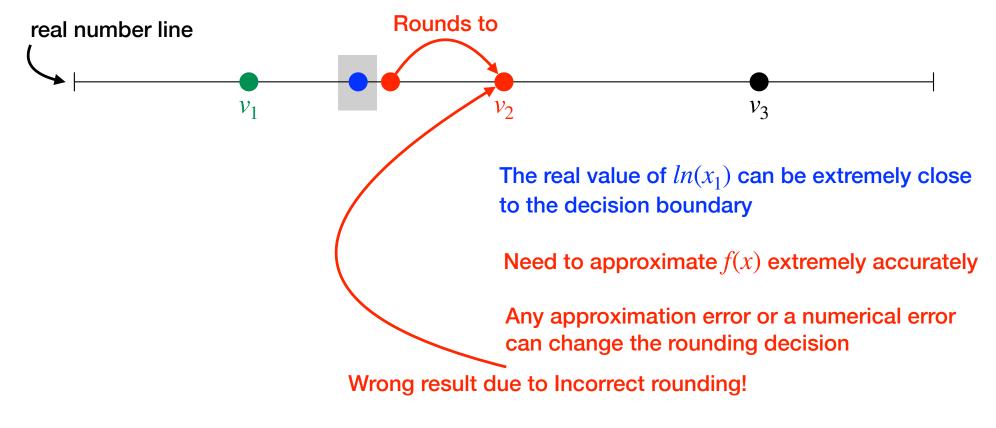


#### What's the issue with Mini-Max Approximations?

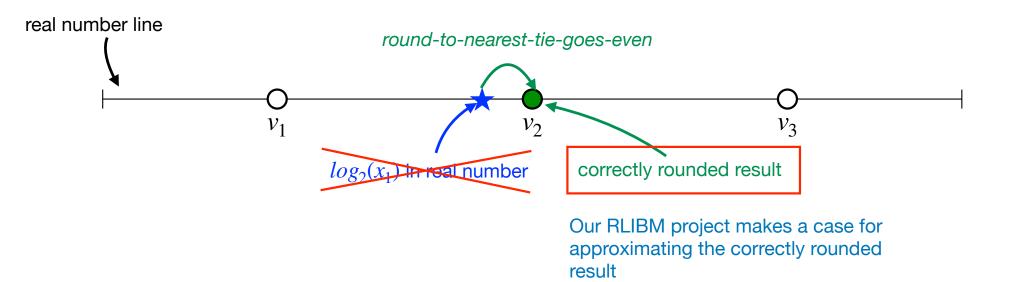


**RUTGERS** 

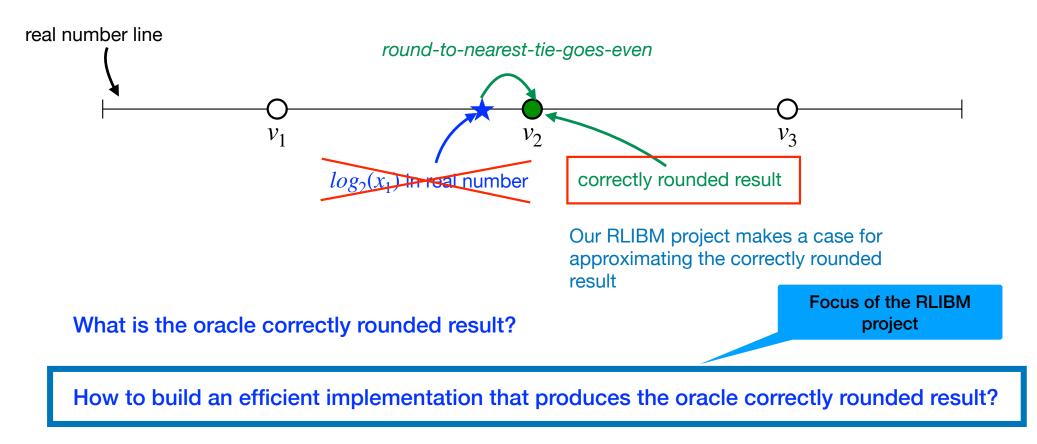
#### What's the issue with Mini-Max Approximations?



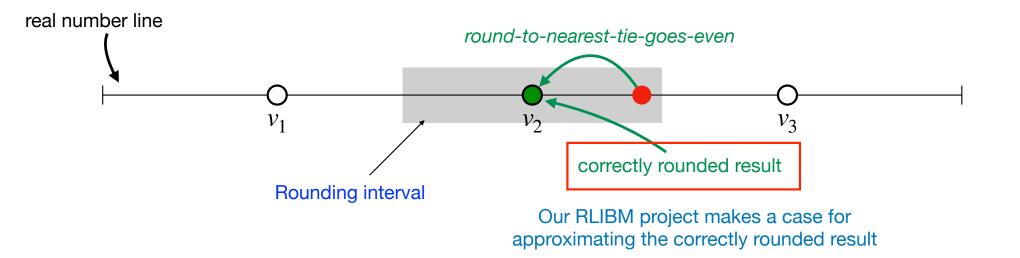






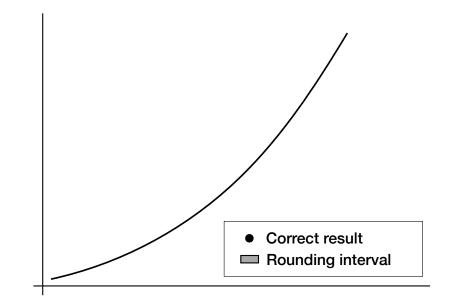


**RUTGERS** 





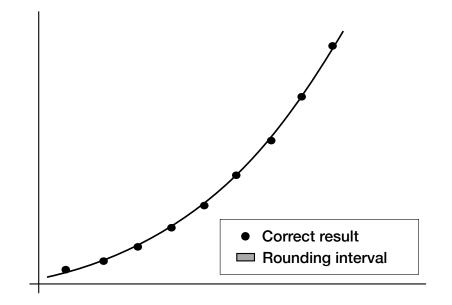
• Given f(x), a representation, and a rounding mode





• Given f(x), a representation, and a rounding mode

1. Compute the correctly rounded result of f(x)

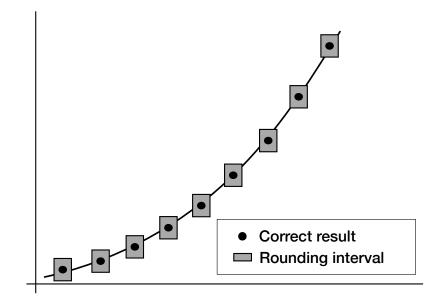




• Given f(x), a representation, and a rounding mode

1. Compute the correctly rounded result of f(x)

2. Identify rounding interval for each input





• Given f(x), a representation, and a rounding mode

1. Compute the correctly rounded result of f(x)

- 2. Identify rounding interval for each input
  - A linear constraint on the output of the polynomial

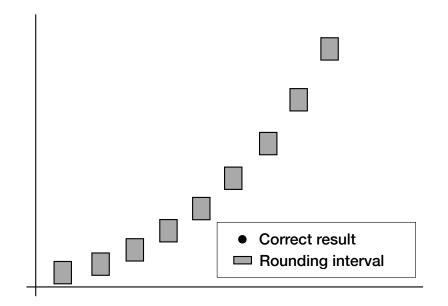
$$l_1 \le P(x_1) \le h_1$$
  

$$l_2 \le P(x_2) \le h_2$$
  

$$l_3 \le P(x_3) \le h_3$$
  

$$l_4 \le P(x_4) \le h_4$$

. . .





• Given f(x), a representation, and a rounding mode

1. Compute the correctly rounded result of f(x)

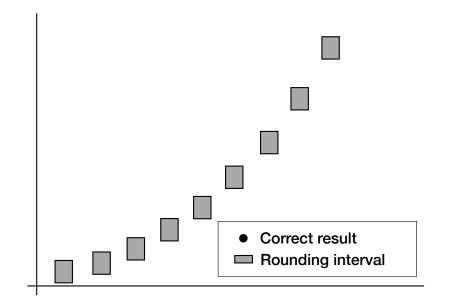
2. Identify rounding interval for each input

• A linear constraint on the output of the polynomial

3. Encode constraint into system of linear inequalities

$$\begin{bmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ \cdots \end{bmatrix} \leq \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ 1 & x_3 & x_3^2 & \cdots & x_3^d \\ 1 & x_4 & x_4^2 & \cdots & x_4^d \\ \cdots \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \cdots \\ c_d \end{bmatrix} \leq \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ \cdots \end{bmatrix}$$

RUTGERS



22

• Given f(x), a representation, and a rounding mode

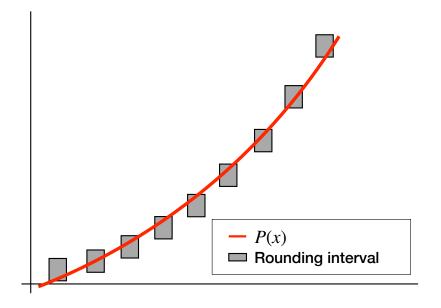
1. Compute the correctly rounded result of f(x)

2. Identify rounding interval for each input

- A linear constraint on the output of the polynomial
- 3. Encode constraint into system of linear inequalities

4. Use a Linear Programming solver to solve for P(x)

$$\begin{bmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ \dots \end{bmatrix} \leq \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ 1 & x_3 & x_3^2 & \dots & x_3^d \\ 1 & x_4 & x_4^2 & \dots & x_4^d \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \dots \\ c_d \end{bmatrix} \leq \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ \dots \end{bmatrix}$$



23

• Given f(x), a representation, and a rounding mode

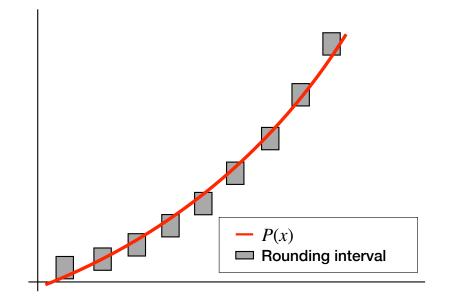
1. Compute the correctly rounded result of f(x)

2. Identify rounding interval for each input

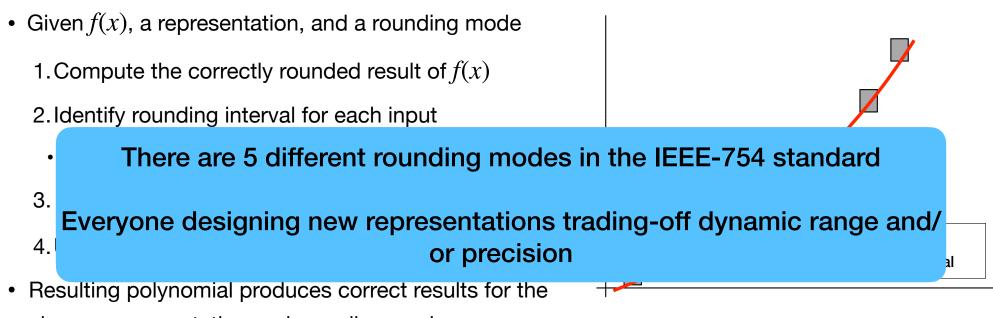
- A linear constraint on the output of the polynomial
- 3. Encode constraint into system of linear inequalities

4. Use Linear Programming solver to solve for P(x)

• Resulting polynomial produces correct results for the chosen representation and rounding mode







chosen representation and rounding mode



#### How do we produce a single polynomial approximation that produces correctly rounded results for multiple representations and rounding modes?





Let's say we want to produce correctly rounded result of  $log_2(x)$  for a 32-bit float

 $log_2(x)$  for 64-bit double type

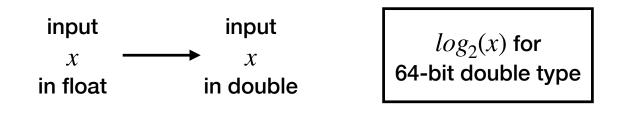


Let's say we want to produce correctly rounded result of  $log_2(x)$  for a 32-bit float

input

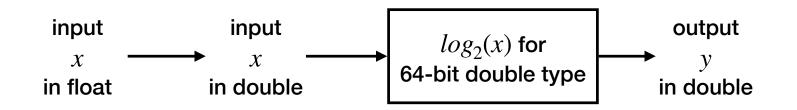
*x* in float  $log_2(x)$  for 64-bit double type



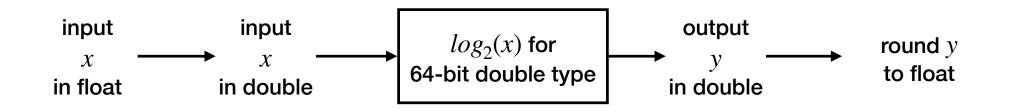




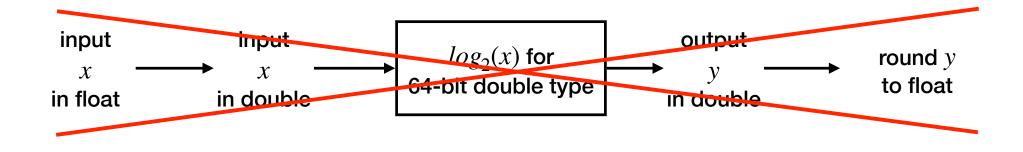
Let's say we want to produce correctly rounded result of  $log_2(x)$  for a 32-bit float



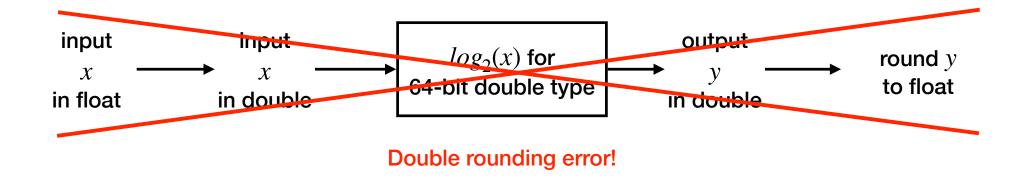








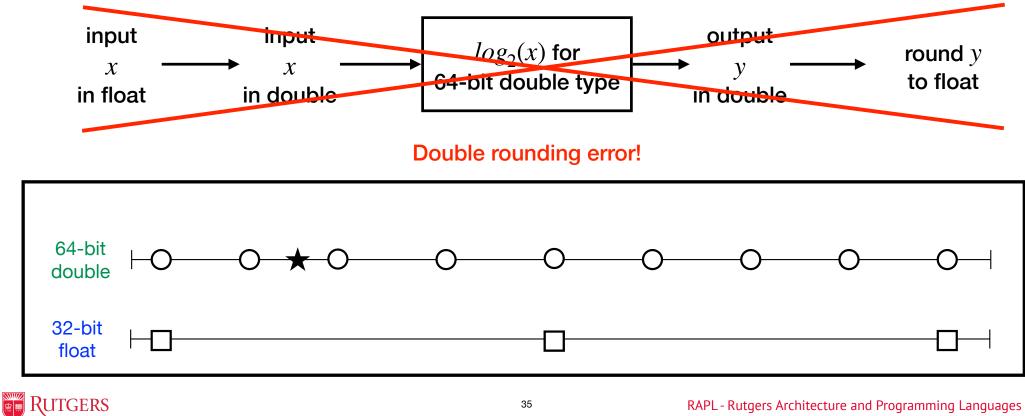






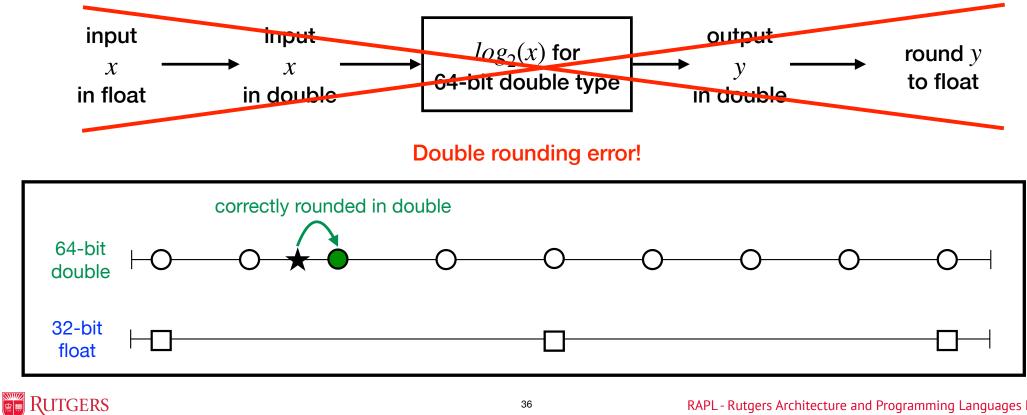
#### **Double Rounding Is The Enemy**

Let's say we want to produce correctly rounded result of  $log_2(x)$  for a 32-bit float



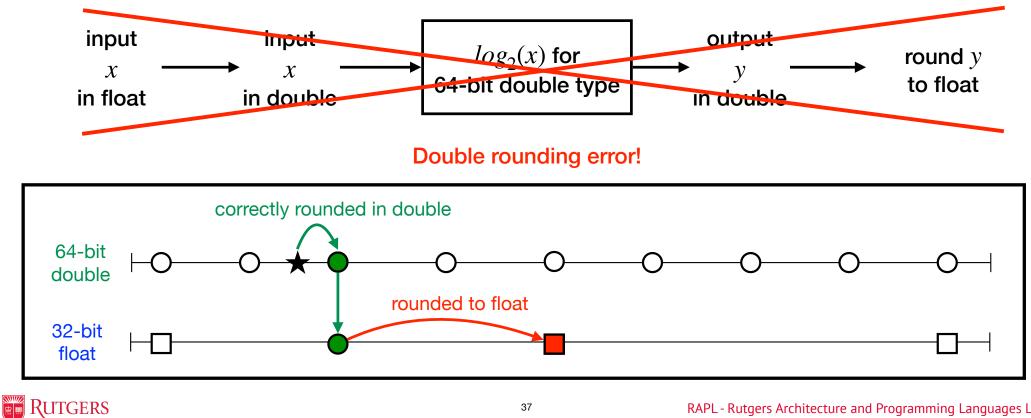
#### **Double Rounding Is The Enemy**

Let's say we want to produce correctly rounded result of  $log_2(x)$  for a 32-bit float



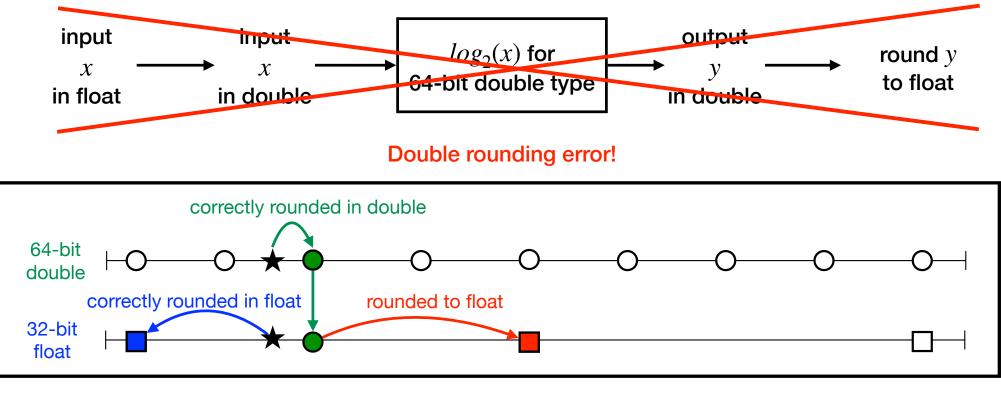
### **Double Rounding Is The Enemy**

Let's say we want to produce correctly rounded result of  $log_2(x)$  for a 32-bit float



### **Double Rounding Is The Enemy**

Let's say we want to produce correctly rounded result of  $log_2(x)$  for a 32-bit float



38

**RUTGERS** 

- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representations
  - Generate a polynomial for the 34-bit FP
  - Using the round-to-odd rounding mode

Generalize: (n + 2) bit floating point representation

### **Round-to-Odd Rounding Mode**

- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP
  - Using the round-to-odd rounding mode
- Round-to-odd:
  - Used for rounding from a decimal to a binary fraction [Goldberg 1991]
  - Used for primitive operations in extended precision [Boldo et al. 2005]



Generalize: (n + 2) bit floating point representation

### **Round-to-Odd Rounding Mode**

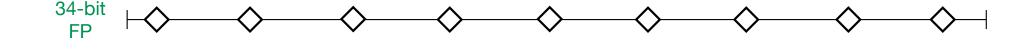
- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP
  - Using the round-to-odd rounding mode
- Round-to-odd:
  - Used for rounding from a decimal to a binary fraction [Goldberg 1991]
  - Used for primitive operations in extended precision [Boldo et al. 2005]
- How do we make it work for elementary functions?
  - Extremely challenging using prior approaches (e.g., Remez Algorithm)
    - How to perform error analysis when round-to-odd mode is involved?
  - Our **RLIBM** approach allows straight-forward integration with round-to-odd!

Generalize: (n + 2) bit floating point representation

### How Does Round-to-Odd Work?

- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP
  - Using the round-to-odd rounding mode
- Round-to-odd:

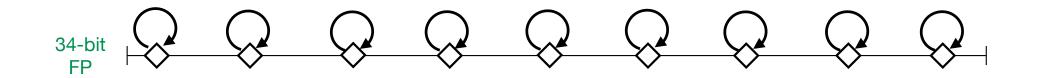
Generalize: (n + 2) bit floating point representation





### How Does Round-to-Odd Work?

- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP
  - Using the round-to-odd rounding mode
- Round-to-odd:
  - If exactly representable, then it is represented with the value

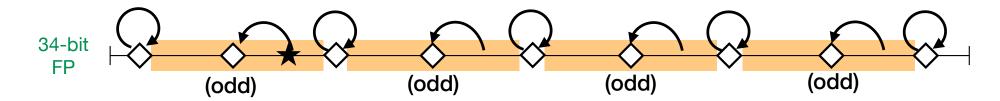




Generalize: (n + 2) bit floating point representation

### How Does Round-to-Odd Work?

- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP
  - Using the round-to-odd rounding mode
- Round-to-odd:
  - If exactly representable, then it is represented with the value
  - Otherwise, rounds to the adjacent odd value



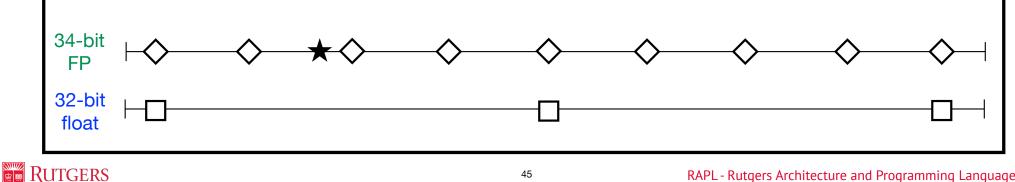


Generalize: (n + 2) bit floating point representation

- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP

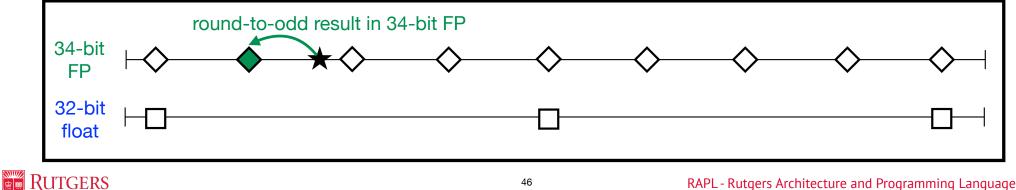
Generalize: (n + 2) bit floating point representation

- Using the round-to-odd rounding mode
- Producing correctly rounded results for a FP type with n-bits or less than n-bit:

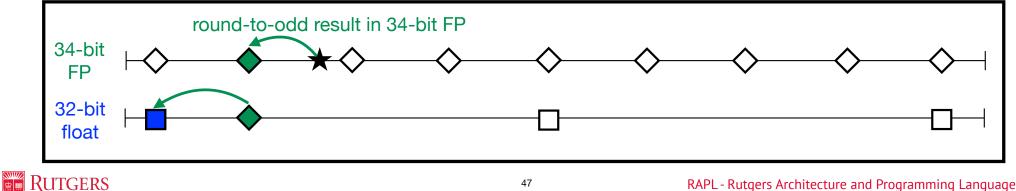


- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP
- Generalize: (n + 2) bit floating point representation

- Using the round-to-odd rounding mode
- Producing correctly rounded results for a FP type with n-bits or less than n-bit:
  - Produce correctly rounded results for (n+2)-bit FP in the round-to-odd mode

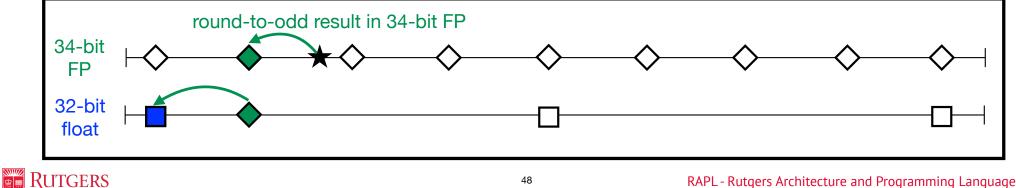


- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP Generalize: (n + 2) bit floating point representation
  - Using the round-to-odd rounding mode
- Producing correctly rounded results for a FP type with n-bits or less than n-bit:
  - Produce correctly rounded results for (n+2)-bit FP in the round-to-odd mode
  - Round the result to FP type with n-bits or less than n-bits using any IEEE-754 rounding mode



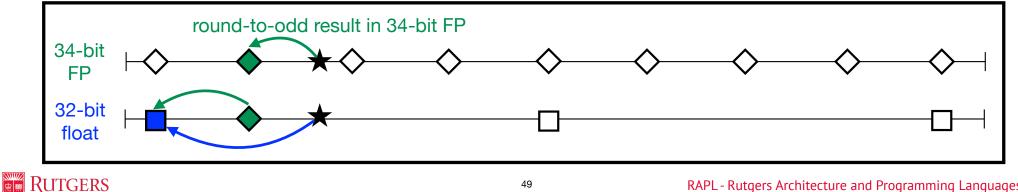
47

- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP Generalize: (n + 2) bit floating point representation
  - Using the round-to-odd rounding mode
- Producing correctly rounded results for a FP type with n-bits or less than n-bit:
  - Produce correctly rounded results for (n+2)-bit FP in the round-to-odd mode
  - Round the result to FP type with n-bits or less than n-bits using any IEEE-754 rounding mode
  - Guaranteed to produce correctly rounded results!

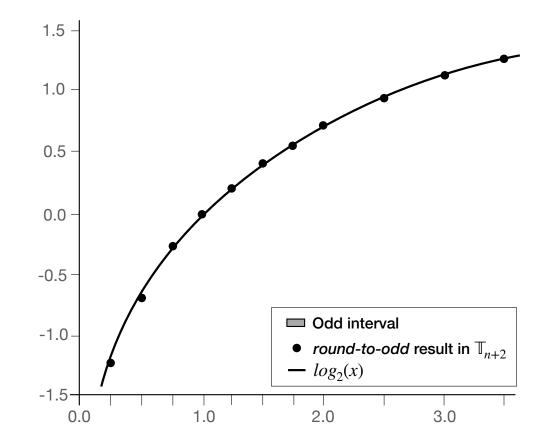


48

- Insight: Retain enough information about the real value even when double rounding
- How to generate 1 polynomial for 10 to 32-bit FP representation:
  - Generate a polynomial for the 34-bit FP Generalize: (n + 2) bit floating point representation
  - Using the round-to-odd rounding mode
- Producing correctly rounded results for a FP type with n-bits or less than n-bit:
  - Produce correctly rounded results for (n+2)-bit FP in the round-to-odd mode
  - Round the result to FP type with n-bits or less than n-bits using any IEEE-754 rounding mode
  - Guaranteed to produce correctly rounded results!



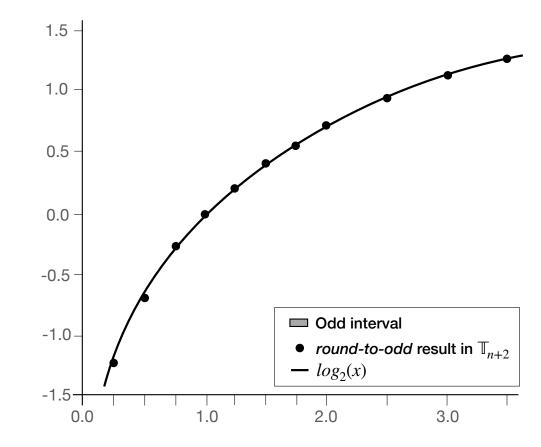
• For each float input, compute





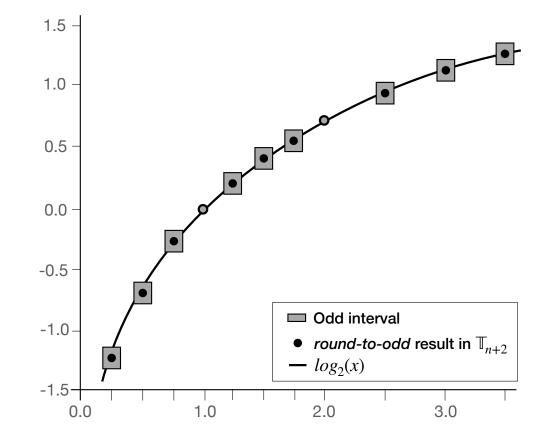


- For each float input, compute
  - the *round-to-odd* result of  $log_2(x)$  in FP34



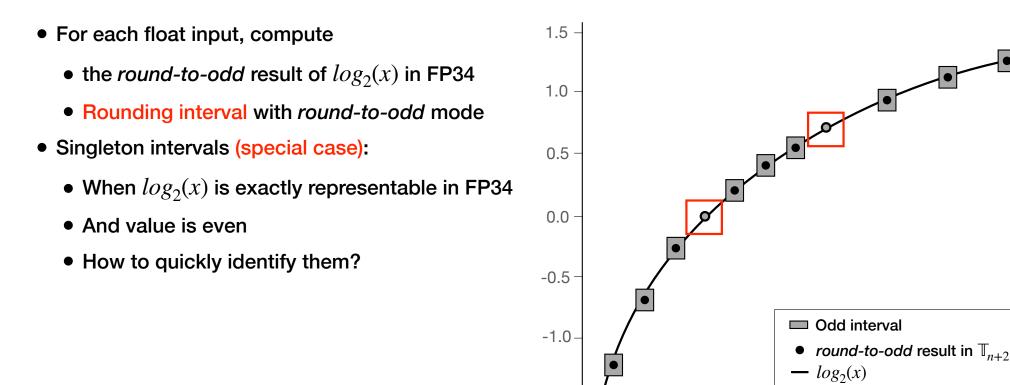


- For each float input, compute
  - the *round-to-odd* result of  $log_2(x)$  in FP34
  - Rounding interval with round-to-odd mode











3.0

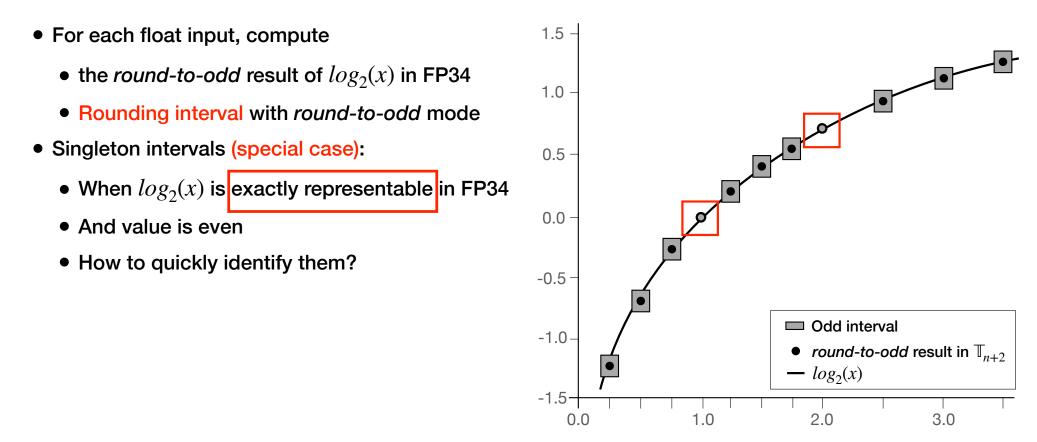
2.0

1.0

-1.5

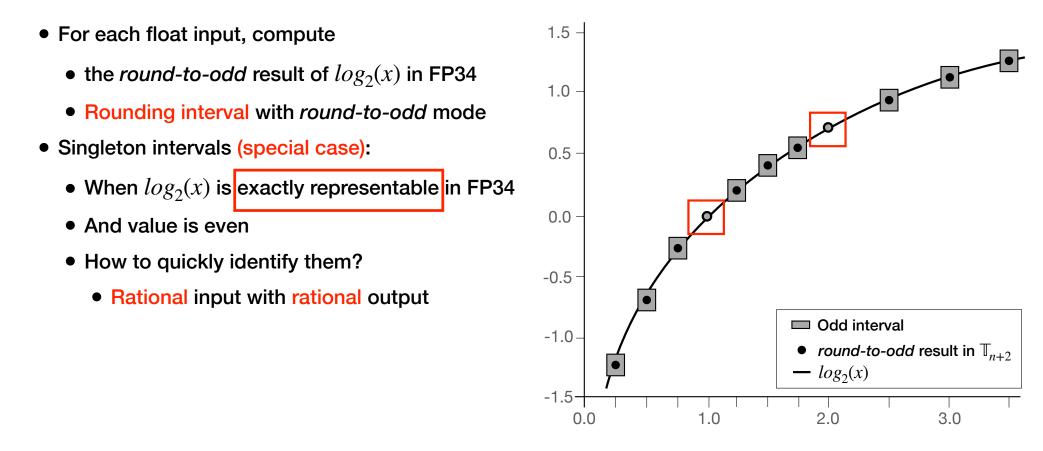
0.0

**RUTGERS** 



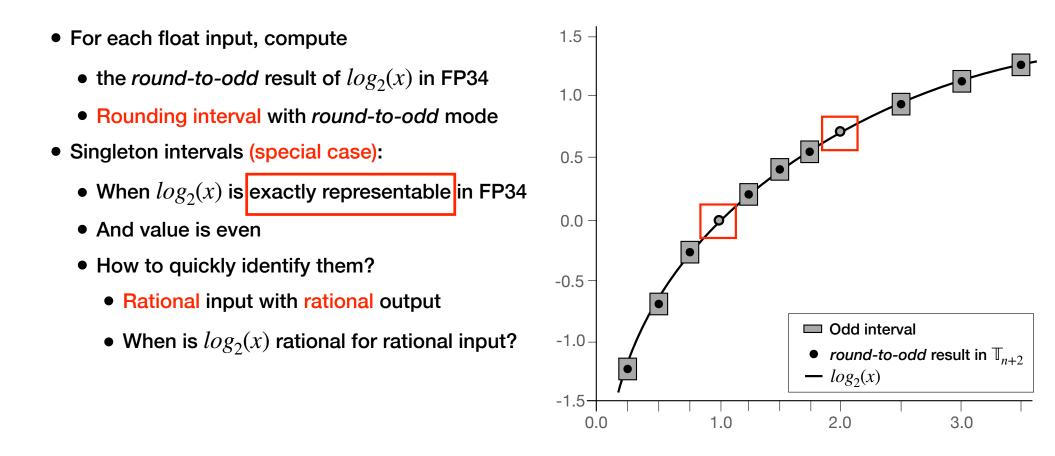
RAPL - Rutgers Architecture and Programming Languages Lab

**RUTGERS** 

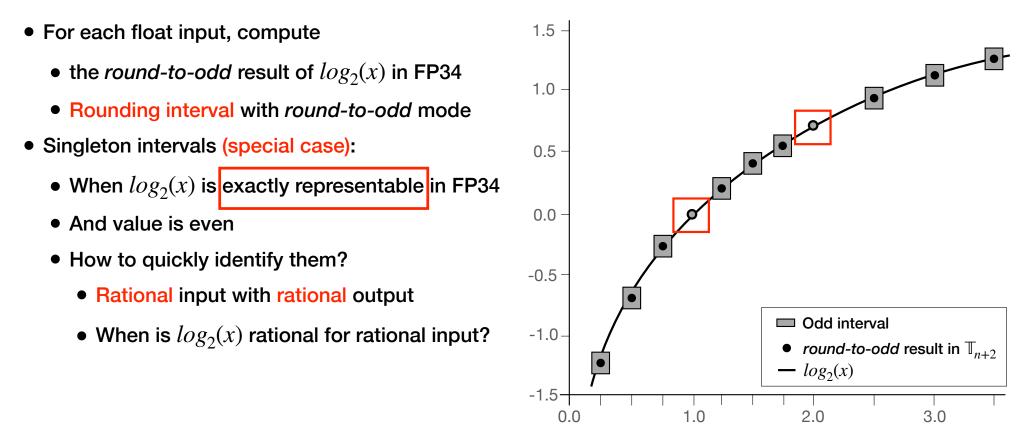


RAPL - Rutgers Architecture and Programming Languages Lab

**RUTGERS** 

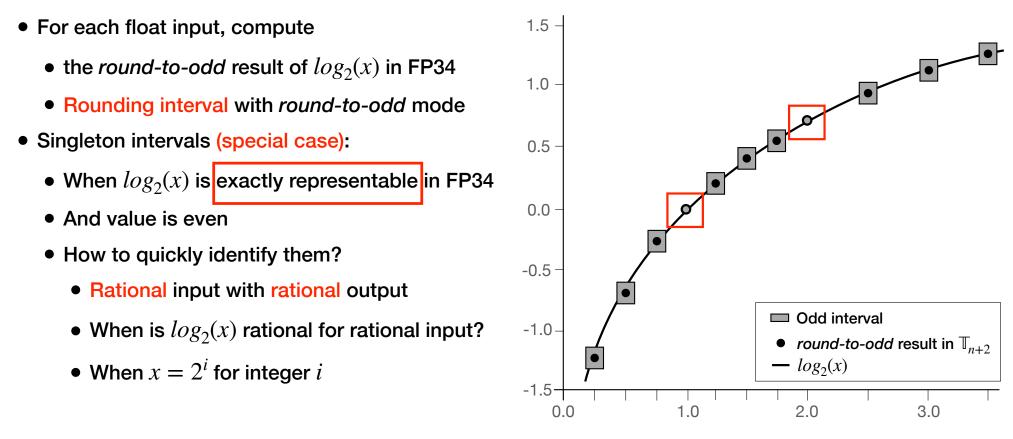


RAPL - Rutgers Architecture and Programming Languages Lab



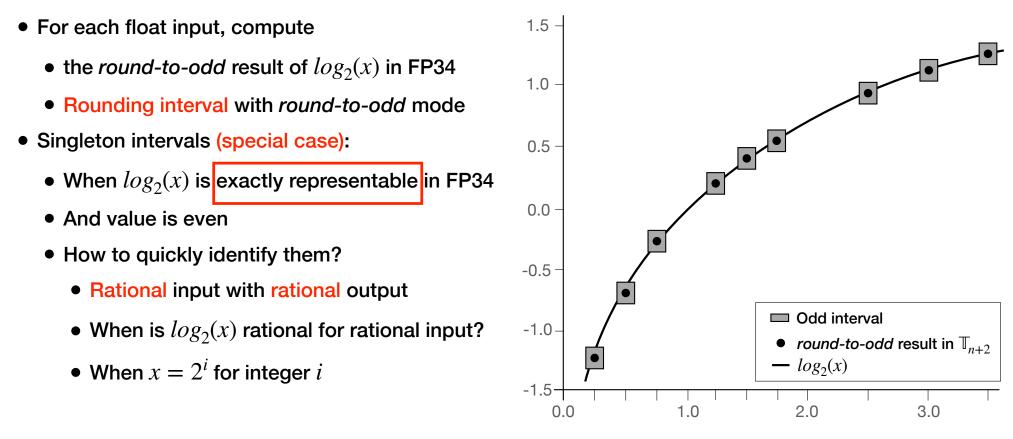
#### Several results regarding rationality of elementary functions!





#### Several results regarding rationality of elementary functions!

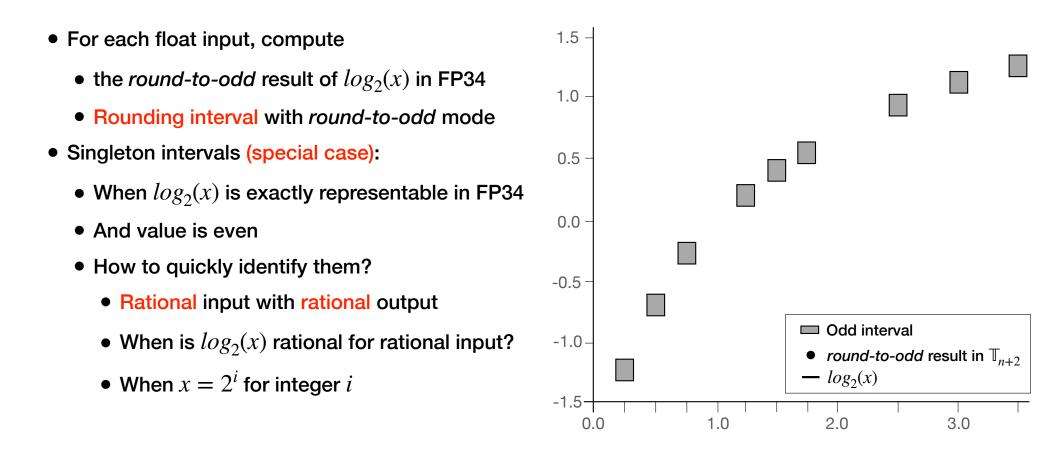
**RUTGERS** 



#### Several results regarding rationality of elementary functions!

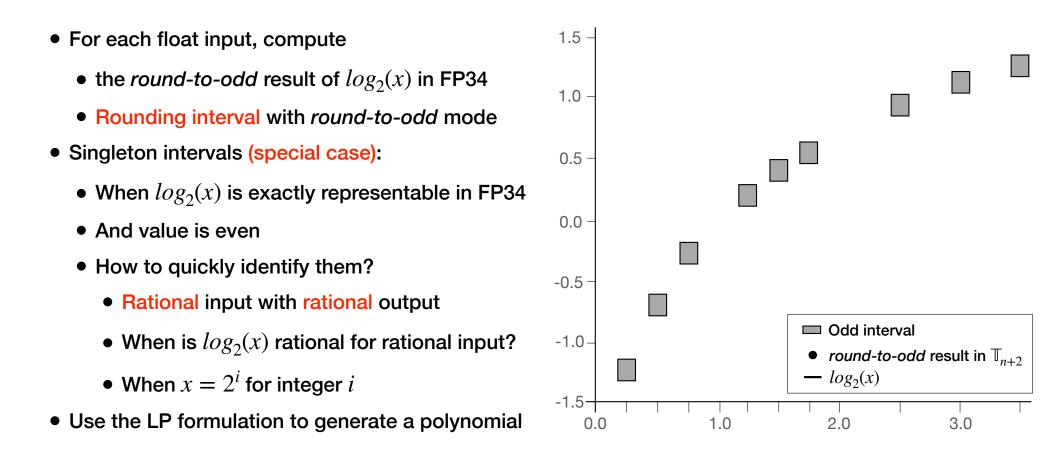
**RUTGERS** 

**RUTGERS** 



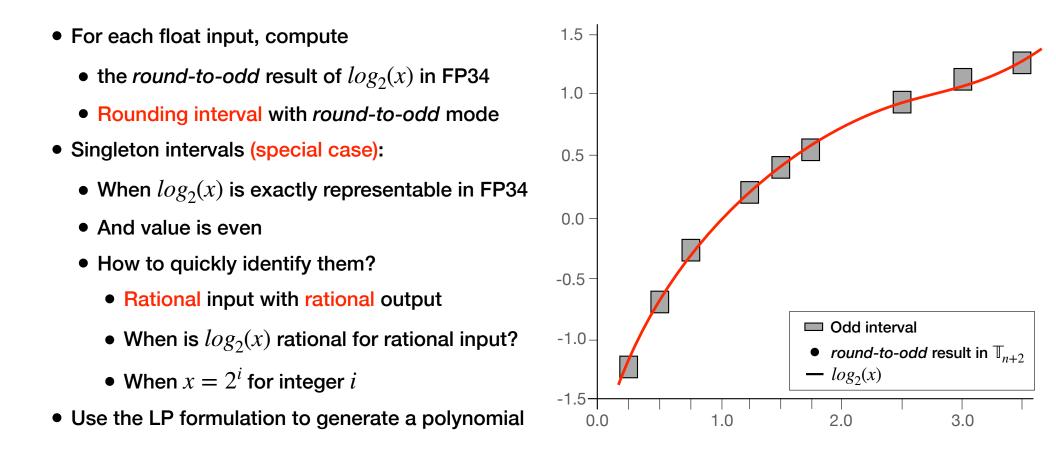
RAPL - Rutgers Architecture and Programming Languages Lab

**RUTGERS** 



RAPL - Rutgers Architecture and Programming Languages Lab

**RUTGERS** 



RAPL - Rutgers Architecture and Programming Languages Lab

### **Does it Work?**



# **Our RLIBM Functions Are Correctly Rounded**

float functions	Using RLIBM-ALL
ln(x)	<ul> <li>✓</li> </ul>
log2(x)	<ul> <li>Image: A start of the start of</li></ul>
log10(x)	<ul> <li>✓</li> </ul>
exp(x)	<ul> <li>✓</li> </ul>
exp2(x)	<ul> <li>✓</li> </ul>
exp10(x)	<ul> <li>✓</li> </ul>
sinh(x)	<ul> <li>✓</li> </ul>
cosh(x)	<ul> <li>✓</li> </ul>
sinpi(x)	<ul> <li>✓</li> </ul>
cospi(x)	<ul> <li></li> </ul>

- Produces correctly rounded results for multiple representations
  - $\leq 8$  exponent bits (Same or less than 32-bit float)
  - $\leq 23$  mantissa bits (same or less than 32-bit float)
  - 161 different configurations
  - Includes float, bfloat16, Tensorfloat32, half, etc
- Supports all five standard rounding modes
- $161 \times 5 = 805$  combinations of configurations and rounding modes



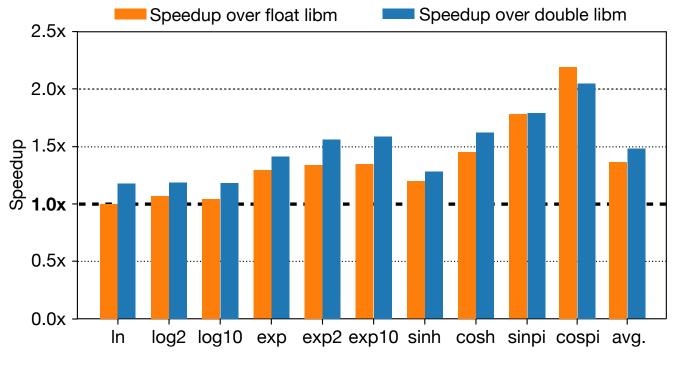
# **Our RLIBM Functions Are Correctly Rounded**

Ability to produce correctly rounded float value with all standard rounding modes for all inputs

float functions	Using RLIBM-ALL	Using glibc (float)	Using glibc (double)	Using Intel (float)	Using Intel (double)	Using CRLibm (double)
ln(x)	<ul> <li>✓</li> </ul>	×	×	×	×	×
log2(x)	<ul> <li>Image: A start of the start of</li></ul>	×	<ul> <li>✓</li> </ul>	×	<ul> <li>✓</li> </ul>	<ul> <li>✓</li> </ul>
log10(x)	<ul> <li>✓</li> </ul>	×	×	×	×	×
exp(x)	<ul> <li>✓</li> </ul>	×	×	×	×	<ul> <li>✓</li> </ul>
exp2(x)	<ul> <li>✓</li> </ul>	×	×	×	×	N/A
exp10(x)	<ul> <li>✓</li> </ul>	×	×	×	×	N/A
sinh(x)	<ul> <li>✓</li> </ul>	×	×	×	×	×
cosh(x)	<ul> <li>✓</li> </ul>	×	×	×	×	<b>v</b>
sinpi(x)	<ul> <li>✓</li> </ul>	N/A	N/A	×	×	<b>v</b>
cospi(x)	<ul> <li>✓</li> </ul>	N/A	N/A	×	×	<ul> <li>✓</li> </ul>



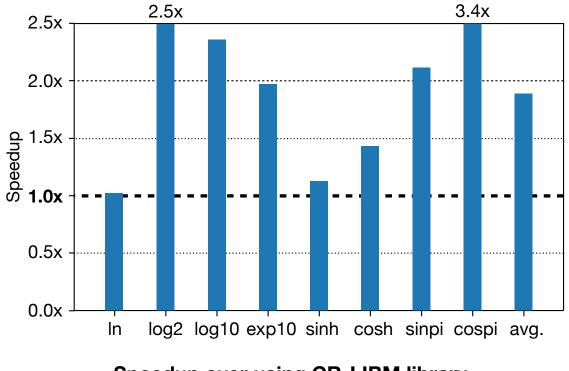
## **RLIBM-ALL Functions Are Fast**



Speedup over using Intel math library



### **RLIBM-ALL Functions Are Fast**



Speedup over using CR-LIBM library



### **Transition to Practice**

### 5 functions in LLVM's libc is built using RLIBM's implementations!

] Implement correctly rounded log2f based on RLIBM library.		Browse files
ement log2f based on RLIBM library correctly rounded for all rounding modes.		
ibc] Implement log10f correctly rounded for all rounding modes.		Browse files
d on RLIBM implementation similar to logf and log2f. Most of the exception	onal inputs are the exact powers of 10.	
[libc] Improve the performance of expf.	Browse files	
Reduce the polynomial's degree from 7 down to 4.		
[libc] Improve the performance of exp2f.	Browse files	
Reduce the range-reduction table size from 128 entries down to 64 entries, and reduce the polynomial's degree from 6 down to 4.		
Currently we use a degree-6 minimax polynomial on an interval of length $2^{-7}$ around 0 to compute exp2f. Based on the suggestion of <b>@santoshn</b> and the RLIBM		
<pre>project (https://github.com/rutgers-apl/rlibm-prog/blob/main/libm/float/exp2.c)</pre>		j109acea92e1acb661c404fa62b9
it is possible to have a good polynomial of degree-4 on a subinterval of length $2^{(-6)}$ to approximate $2^{x}$ .		)109acea92e1acb00104041a02b9
We did try to either reduce the degree of the polynomial down to 3 or increase		
the interval size to $2^{(-5)}$ , but in both cases the number of exceptional values exploded. So we settle with using a degree-4 polynomial of the interval of		
size 2 <sup>(-6)</sup> around 0.		
Reviewed By: michaelrj, sivachandra, zimmermann6, santoshn		
Differential Revision: https://reviews.llvm.org/D122346		

# Conclusion

- Approximate the correctly rounded result rather than the real value
  - Linear programming formulation based on the interval around the correct result
- How to generate a generic polynomial for all representations up to n-bits?
  - Generate polynomials for (n+2)-bit representation
  - That produces correct results with round-to-odd mode
  - Quickly identify singleton intervals: when f(x) is rational for rational input
- Our RLIBM prototype produces correctly rounded results for all inputs in
  - 161 different FP configurations
  - All 5 IEEE-754 standard rounding modes
- Faster than mainstream math libraries

Make correctly rounded results mandatory rather than a recommendation by the standards

### **Open Source**

# Visit the RLIBM page for papers & prototypes https://www.cs.rutgers.edu/~santosh.nagarakatte/rlibm/

Thanks National Science Foundation for the support



