LOW POWER PROBABILISTIC FLOATING POINT MULTIPLIER DESIGN

Aman Gupta[‡,†,*], Satyam Mandavalli[‡], Vincent J. Mooney[$,†,*,§], Keck-Voon Ling[†,*], Arindam Basu[†], Henry Johan[$,*] and Budianto Tandianus[$,*]

[‡]International Institute of Information Technology, Hyderabad, India
[$]School of Computer Engineering, NTU, Singapore
[†]School of EEE, Nanyang Technological University (NTU), Singapore
[*]NTU-Rice Institute of Sustainable and Applied Infodynamics (ISAID), NTU, Singapore
[§]School of ECE, Georgia Institute of Technology, Georgia, USA

Outline

- Motivation and Definition of Probabilistic Computation
- Typical Design
- Low Power Probabilistic Design
  - Probabilistic Floating Point Multiplier
  - Low Power Techniques
  - C simulator
- Ray Tracing Application
- Simulations and Results
- Conclusion

Outline

- **Motivation and Definition of Probabilistic Computation**
- Typical Design
- Low Power Probabilistic Design
  - Probabilistic Floating Point Multiplier
  - Low Power Techniques
  - C simulator
- Ray Tracing Application
- Simulations and Results
- Conclusion

Need for Low Power FP Multiplier

- Wide dynamic number range is provided by floating point format which is used by real time graphics and multimedia applications
- Cost of this dynamic range is a power hungry floating point unit in the architecture
- Fixed power budget systems/devices call for designing of energy efficient floating point units
- Among floating point operations, multiplication has the maximum power consumption in most of the applications
- Hence, the focus of the work presented here is to attain low power floating point multiplication

Probabilistic Computation

- Achieves power savings by trading off the accuracy of the computation
- More significant calculations have a larger contribution in the computed result than less significant calculations
- More energy is invested in more significant calculations and less energy is invested in less significant calculations
- Applications which generate data for human perception can produce “reasonably good” results without requiring exact computations

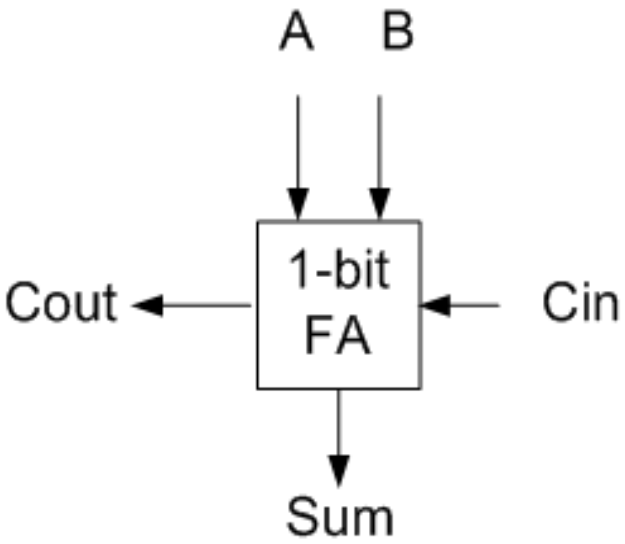
Probabilistic Computation

- Current rate of technology scaling and power supply reduction will make circuits prone to thermal noise due to noise margin reduction
- It is predicted that noise will affect the correct functioning of circuits in future technology nodes*
- International Technology Roadmap for Semiconductors (ITRS) predicts that *relaxing the requirement of 100% correctness for devices and interconnects is likely forced by technology scaling*
- Hence, we model the effect of device noise for future technology nodes in the work presented

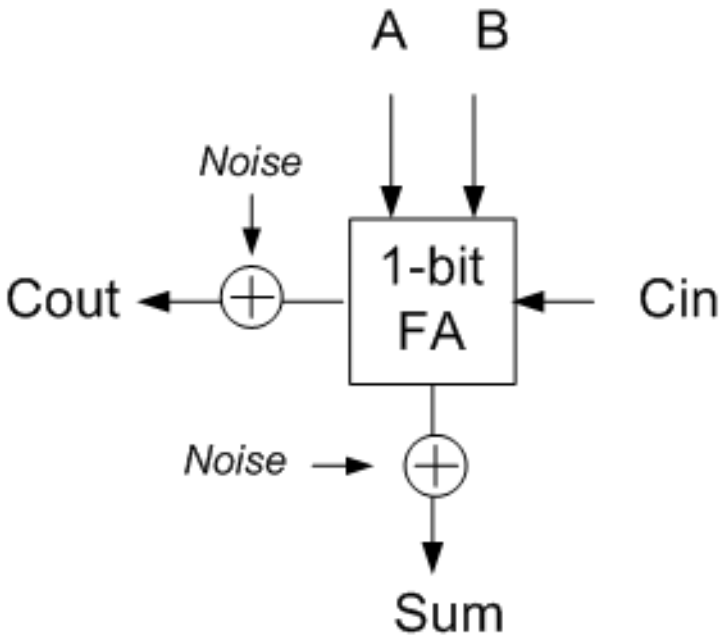
*L. B. Kish, “End of Moore"s law: Thermal (noise) death of integration in micro and nano electronics,” Physics Letters A, 2002, vol. 305, no. 3-4, pp. 144–149.

Noise Modeling

- Deliberate addition of noise sources at the outputs of gates to model noise



Ideal Gate



Noisy Gate

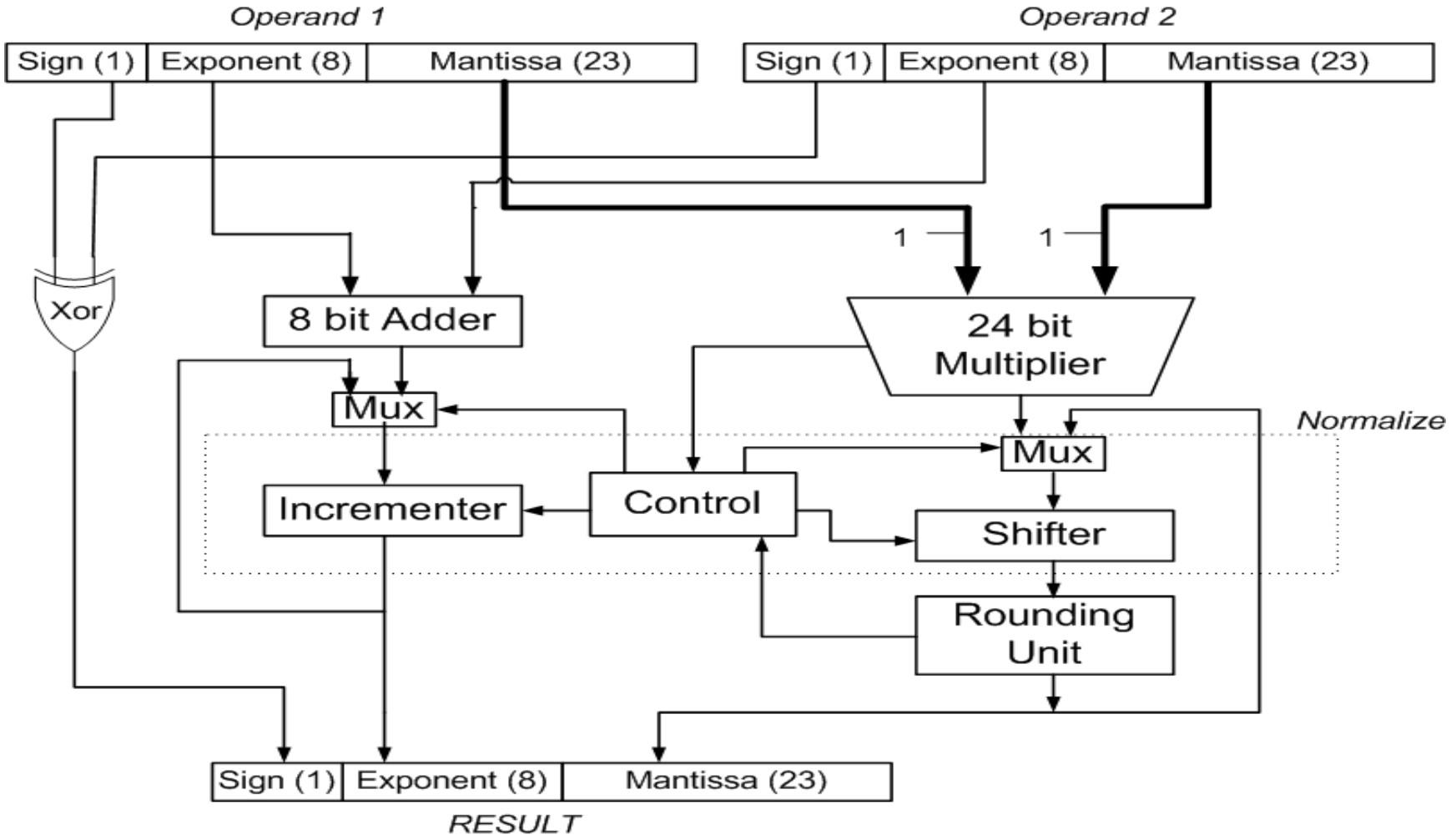
Outline

- Motivation and Definition of Probabilistic Computation
- **Typical Design**
- Low Power Probabilistic Design
  - Probabilistic Floating Point Multiplier
  - Low Power Techniques
  - C simulator
- Ray Tracing Application
- Simulations and Results
- Conclusion

Typical Floating Point Multiplier

- A single precision floating point number(32 bits) has three components, namely,
  - sign(1 bit)
  - exponent(8 bits) and
  - mantissa(23 bits + 1 bit)
- Multiplication of two floating point numbers requires three operations:-
  - Multiplication of the mantissas of the operands
  - Addition of the exponents of the operands and
  - Calculation of the sign bit of the result

Typical Floating Point Multiplier

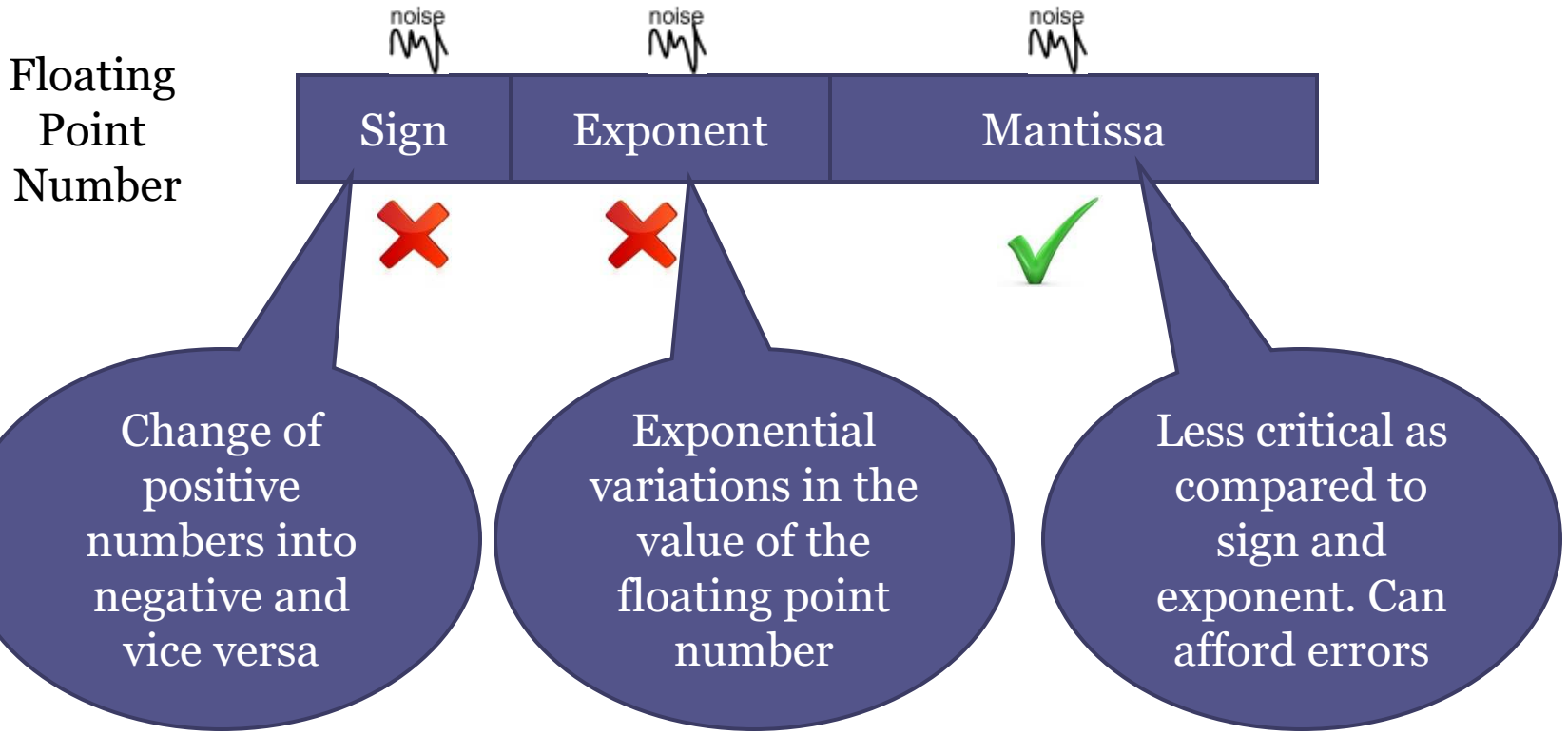


Outline

- Motivation and Definition of Probabilistic Computation
- Typical Design
- **Low Power Probabilistic Design**
  - **Probabilistic Floating Point Multiplier**
  - Low Power Techniques
  - C simulator
- Ray Tracing Application
- Simulations and Results
- Conclusion

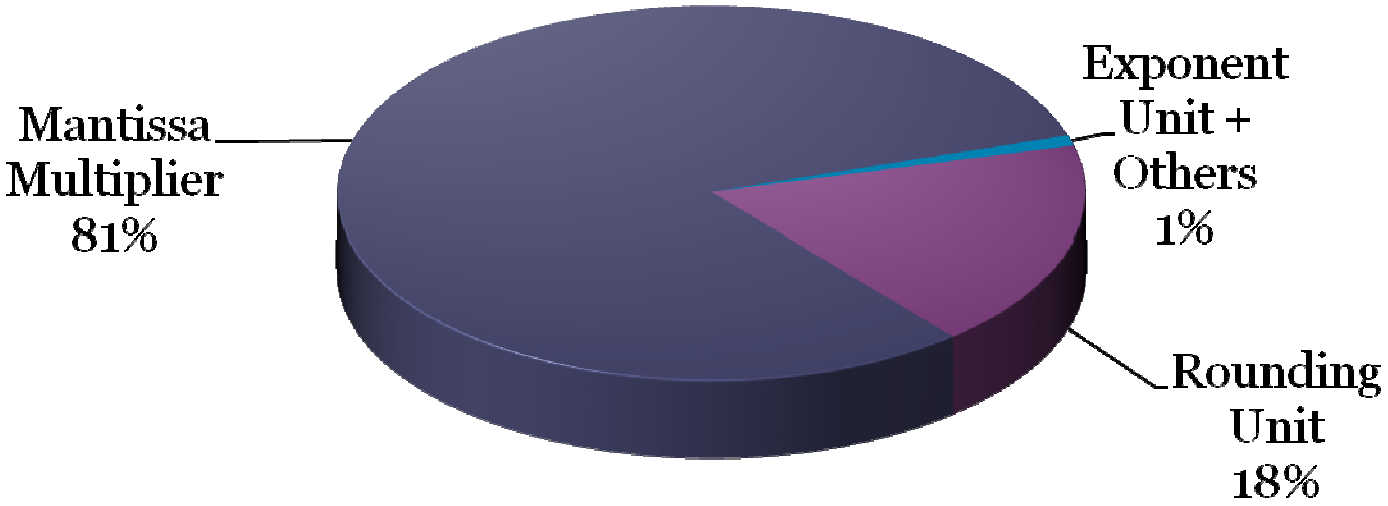
Factors Involved in the Choice of Probabilistic Components

1. Significance of the Calculation



Factors Involved in the Choice of Probabilistic Components

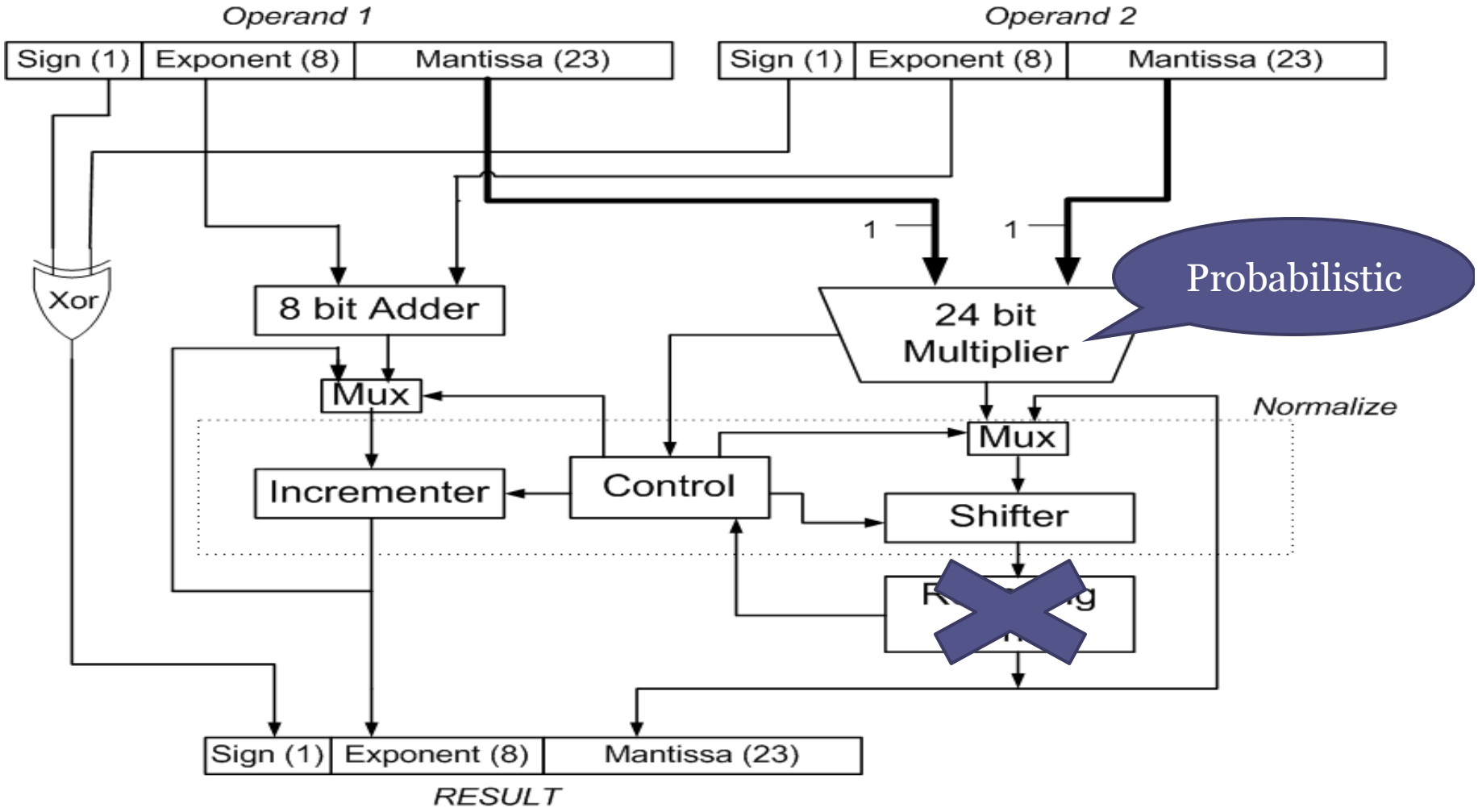
2. Power Consumption of the Computational Blocks
   - Energy savings on the mantissa multiplier block will be substantial energy savings on the overall floating point multiplier
   - As the computation is probabilistic, rounding or not rounding the result does not make any significant change in the accuracy of the results



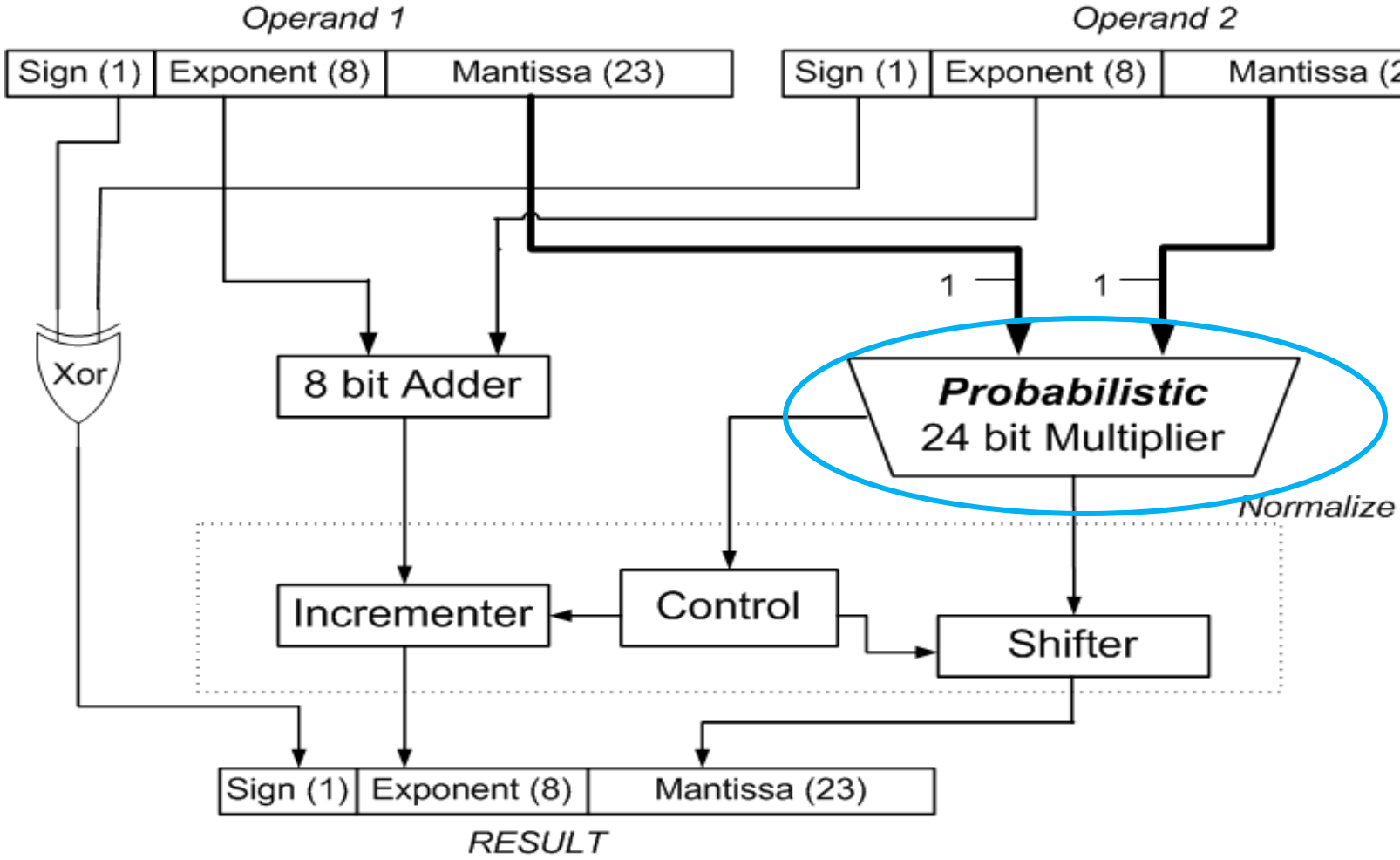
Probabilistic Floating Point Multiplier

- The difference between the typical design and the probabilistic design is that
  - Mantissa calculation is made probabilistic
    - to have minimal effect on the value of the result and
    - to attain maximum energy savings on the overall design
  - Rounding unit is not used

Typical Floating Point Multiplier



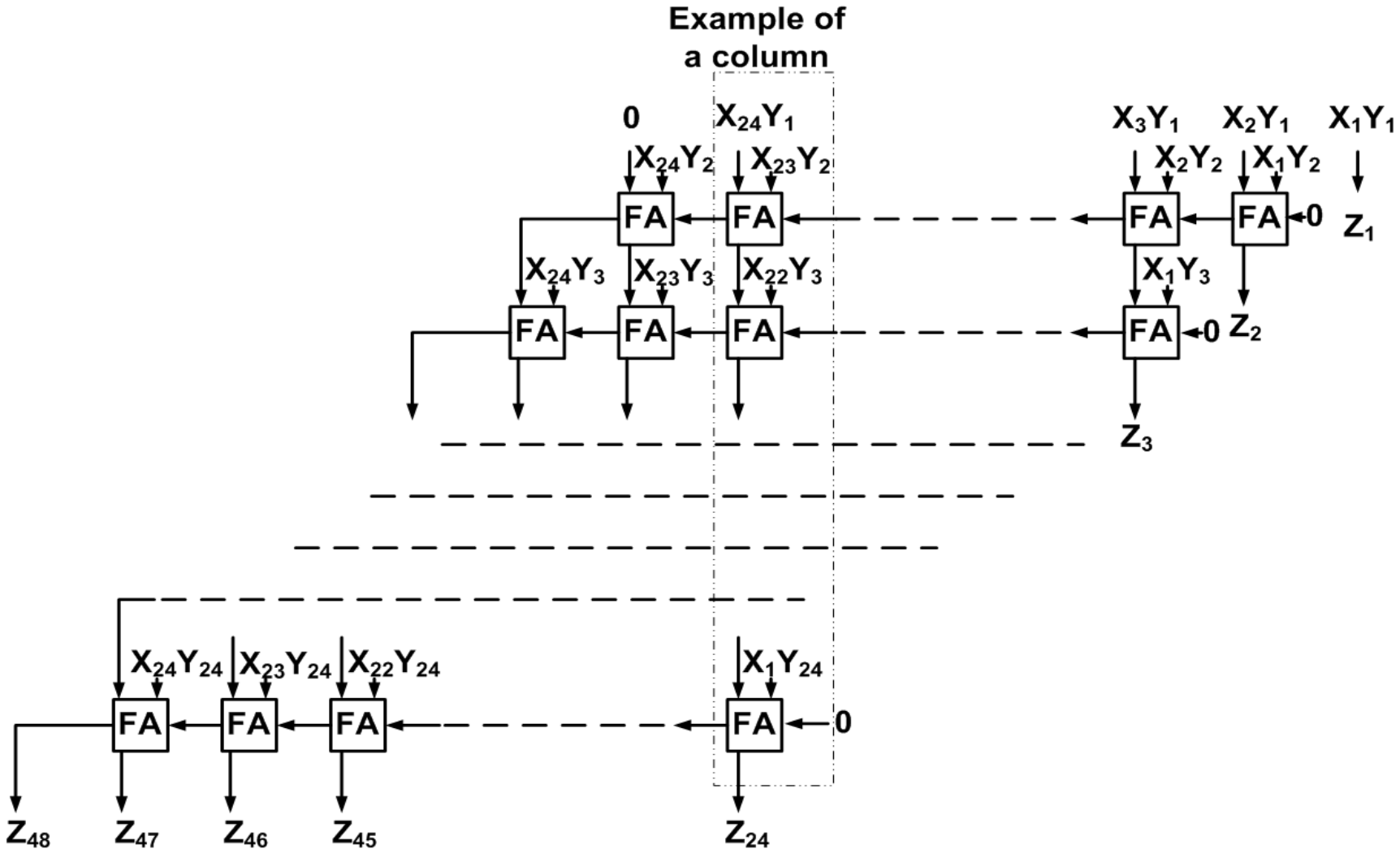
Probabilistic Floating Point Multiplier



Probabilistic 24-bit Mantissa Multiplier

- Array multiplier, the most fundamental multiplier design, is chosen to implement the mantissa multiplication
- The structure of an Array multiplier can be seen as full adders arranged in columns, each column leading to a significant bit of the result
- The full adder columns leading to calculation of more significant bits are termed as more significant columns
- The total number of columns = 46

24-bit Array Multiplier Structure



Outline

- Motivation and Definition of Probabilistic Computation
- Typical Design
- **Low Power Probabilistic Design**
  - Probabilistic Floating Point Multiplier
  - **Low Power Techniques**
  - C simulator
- Ray Tracing Application
- Simulations and Results
- Conclusion

Low Power Techniques Used in the Floating Point Multiplier

- The low power techniques are applied only to the 24-bit mantissa multiplier block
- We use two techniques to attain low power
  - Reduction of the supply voltage of gates, i.e., voltage scaling
  - Putting off/Truncation of some of the gates
- Experiments are performed using Synopsys 90nm library with nominal voltage of 1.2v
- The supply voltage is scaled to following five voltages, namely, 1.2v,1.1v,1.0v,0.9v,0.8v
- A sequence of voltages which defines the supply voltage of each full adder column is termed as a “voltage profile” of a multiplier

Truncation/Sleep

- Some of the less significant full adder columns are truncated
- The remaining full adder columns are operated at the nominal technology voltage
- Example voltage profile for Truncation scheme

Supply Voltage	Truncation (0V)	1.2V
Columns (LSB-MSB)	1-23	24-46

Biased VOltage Scaling (BIVOS)

- Gates of less significance are operated at a lower voltage than the gates of higher significance
- Energy investment is biased to the significance of the calculation
- Full adders in the same column receive the same supply voltage
- Example voltage profile for BIVOS scheme

Supply Voltage	0.8V	0.9V	1.0V	1.1V	1.2V
Columns (LSB-MSB)	1-20	21-29	30-33	34-35	36-46

Proposed scheme BIVOS + Truncation

- Starting from the least significant column
  - some columns are truncated while
  - the rest of the columns are given a biased supply voltage
- The less significant columns, which are not truncated, are operated at a lower voltage as compared to the more significant columns
- Example voltage profile for BIVOS + Truncation scheme

Supply Voltage	Truncation (0V)	0.8V	0.9V	1.0V	1.1V	1.2V
Columns (LSB-MSB)	1-22	23-24	25-29	30-33	34-35	36-46

Outline

- Motivation and Definition of Probabilistic Computation
- Typical Design
- **Low Power Probabilistic Design**
  - Probabilistic Floating Point Multiplier
  - Low Power Techniques
  - **C simulator**
- Ray Tracing Application
- Simulations and Results
- Conclusion

C Simulator

- Simulation of 24-bit Array multiplier takes unreasonable amount of time in HSPICE (300 samples in 24hrs)
- Voltage scaling and thermal noise introduction not feasible in digital simulators
- We developed a simulator in programming language C to calculate
  - The error rate at the output bits and
  - The total energy consumed by the 24-bit Array multiplier used for mantissa multiplication

C Simulator – Error rate calculation

- 1-bit noisy full adder is simulated in HSPICE and error rate is calculated for each voltage*
- An array multiplier model is developed in C using full adders
- Each full adder introduces errors at its outputs in accordance with the supply voltage by using the probability of error values
- The error rate of the output bits is calculated
  - Given a voltage profile and
  - the error rate values from HSPICE

*A. Singh, A. Basu, K.V. Ling and V. J. Mooney, "Modeling Multi-output Filtering Effects in PCMOS," Proceedings of the VLSI Design and Test Conference (VLSI-DAT 2011), April 2011

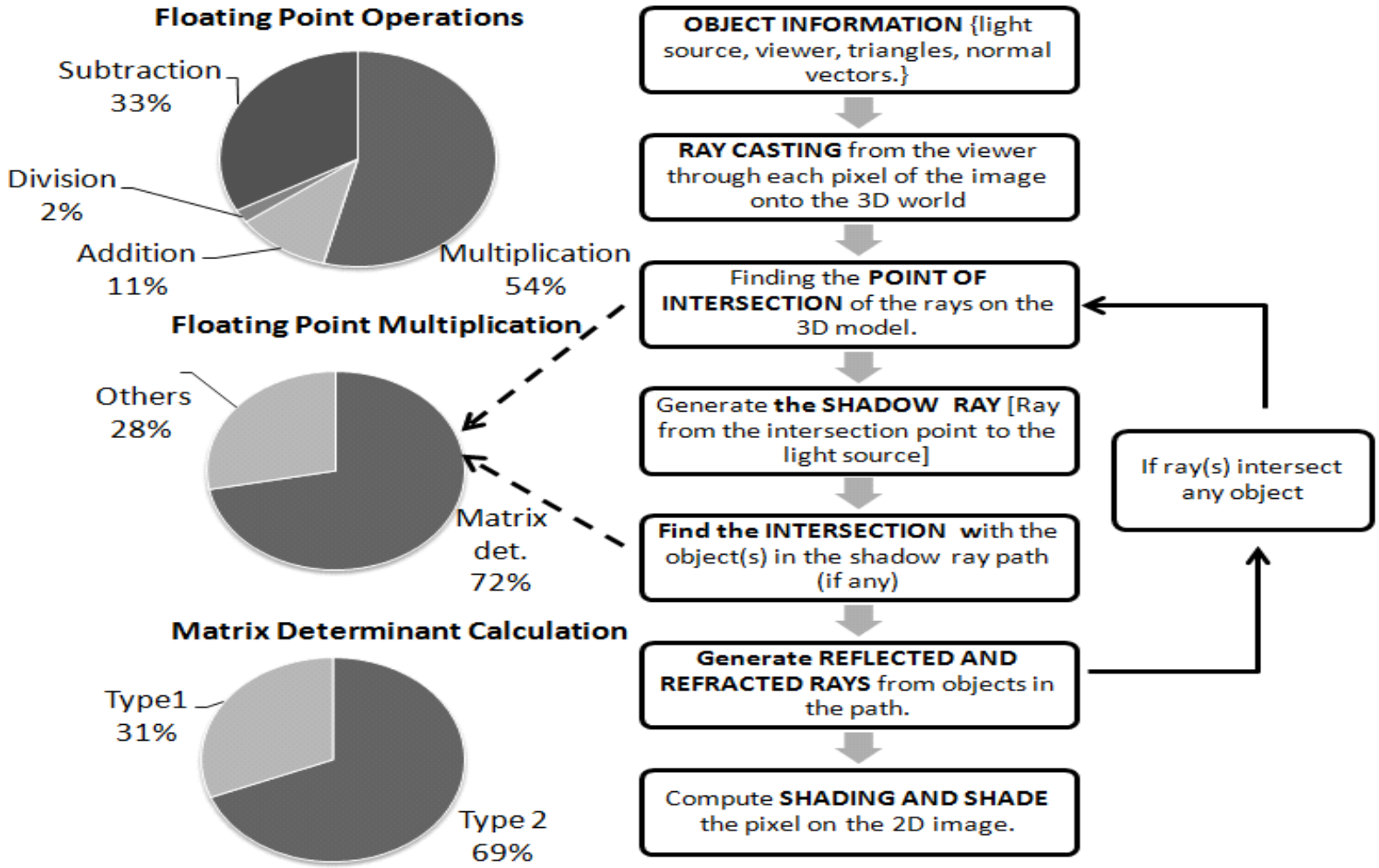
C Simulator – Energy Calculation

- 1-bit probabilistic full adder is simulated in HSPICE for each voltage to calculate
  - Energy per sum toggle
  - Energy per carry toggle
- The total number of toggles for sum and carry for each full adder is calculated using Verilog description in Active HDL software
- Toggle rates multiplied with energy per toggle gives total power consumption

Outline

- Motivation and Definition of Probabilistic Computation
- Typical Design
- Low Power Probabilistic Design
  - Probabilistic Floating Point Multiplier
  - Low Power Techniques
  - C simulator
- **Ray Tracing Application**
- Simulations and Results
- Conclusion

Application Example - Ray Tracing



Application Example - Ray Tracing

- Around 50% of the total multiplications can be made probabilistic without increasing the total number of operations
- The probability of error values generated from the C simulator are used to introduce errors in the floating point multiplication result

Outline

- Motivation and Definition of Probabilistic Computation
- Typical Design
- Low Power Probabilistic Design
  - Probabilistic Floating Point Multiplier
  - Low Power Techniques
  - C simulator
- Ray Tracing Application
- **Simulations and Results**
- Conclusion

Simulation

- Deliberate injection of Gaussian noise at each full adder output to model thermal noise
- Noise RMS chosen such that there are no errors at the output of the full adder when operated at 1.2V
- If supply voltage is lower than 1.2V, there are errors at the output which increase as the voltage is lowered
- This is how we predict a possible noisy future technology node
- This leads to generation of tradeoffs between energy and accuracy as lower energy means lowering of supply voltage which in turn leads to more errors

Simulation

- *Ideal case*
  - no errors in the multiplications of the ray tracing algorithm
  - all the generated images compared with the ideal image
- *Base case*
  - only rounding errors in the multiplications
  - The energy consumed (all FAs at 1.2V) is considered to be 100% energy

Results

PSNR(dB) Relative to the Ideal Image	**Energy Relative to the Base Case**		
	BIVOS	***Truncation/ Sleep***	***BIVOS + Truncation***
58 dB	80%	75%	66%
47 dB	55%	51%	38%

Images with Ray Tracing

Ideal Images

BIVOS + Truncation at 38% Energy

Outline

- Motivation and Definition of Probabilistic Computation
- Typical Design
- Low Power Probabilistic Design
  - Probabilistic Floating Point Multiplier
  - Low Power Techniques
  - C simulator
- Ray Tracing Application
- Simulations and Results
- **Conclusion**

Conclusion

- Energy savings of around 62% can be achieved in a floating point multiplier by using the proposed BIVOS + Truncation scheme
- Overall application level savings = 31% (as 50% of the total multiplications are probabilistic)
- Hence, substantial energy savings can be achieved in floating point multipliers by using probabilistic computing
- Probabilistic computation can be used to harness energy savings in other applications which generate data for human perception

Thank You