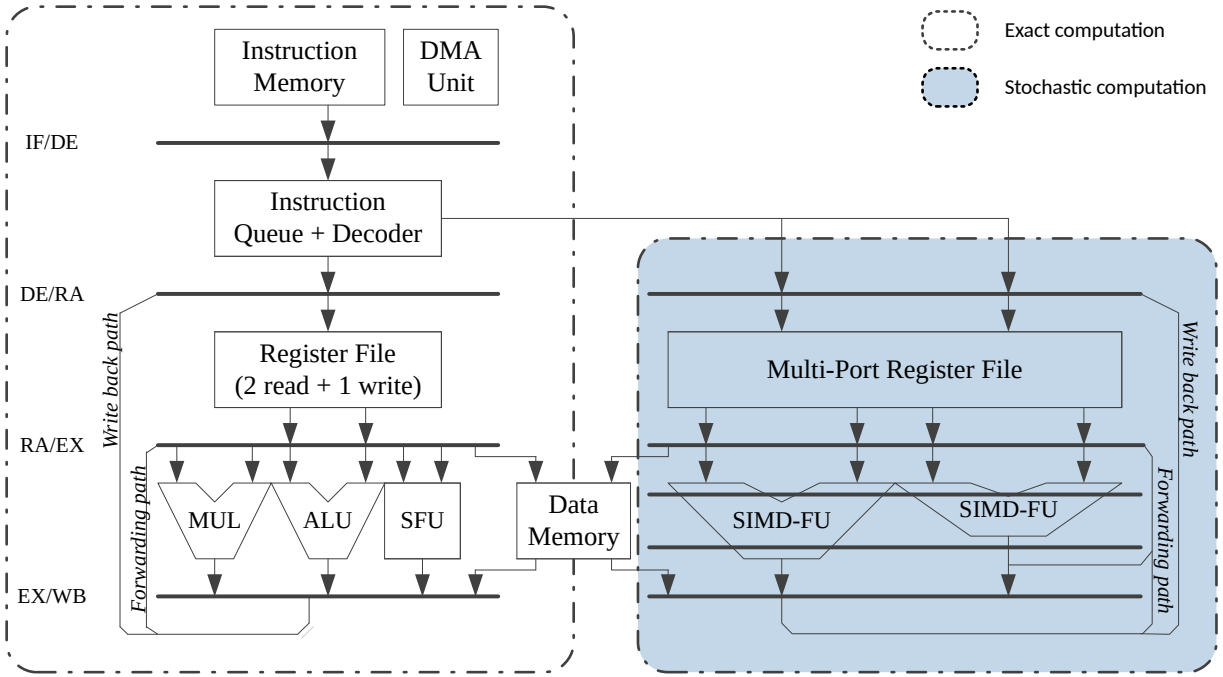
Approximate and stochastic computing: opportunities for processing hardware architectures

Prof. Alberto Garcia-Ortiz

Tensilica Day, Hannover 2019

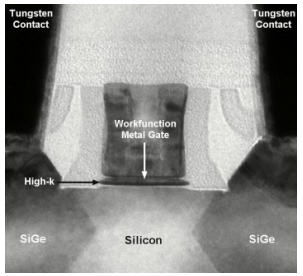
Goal

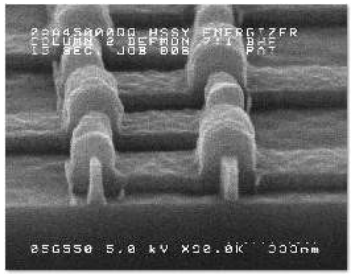
- Opportunities for processing hardware architectures
- Using (imprecise) approximate and stochastic computing



Key idea

- My technology is getting unreliable. What can I do?





- My application can tolerate errors. How can I exploit it?

Goal: Make errors first class citizens

Approximate and stochastic computing: opportunities for processing hardware architectures

- Introduction and motivation
- Approximate processing
- Stochastic processing

Challenges: power crisis

- Low-power = key product differentiator
- Autonomy (energy harvesting, ubiquitous)

- Leakage increasing exponentially

- Thermal issues, including variability
- Packaging cost
  - +8% power, package cost x3
- Reliability
  - +10 °C => +50% failure probability
- Performance drop
  - +10 °C => -3% performance

TZI: Mobile Solutions

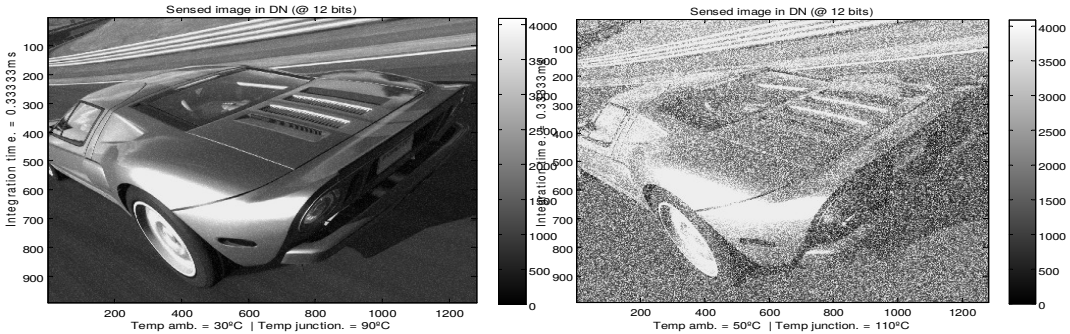


Image degradation in VSoC because of +20C temperature increase

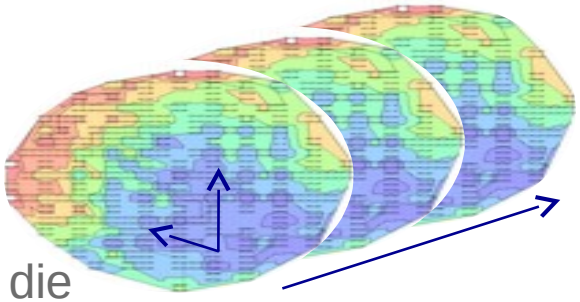
- Environmental concerns: **The best energy source is the energy-save**
  - Data centers consume on the order of 1% of all electricity in 2005
  - On-chip interconnect power exceeds the total solar energy in 2007 (USA)

IT2Green.de

Power-aware design

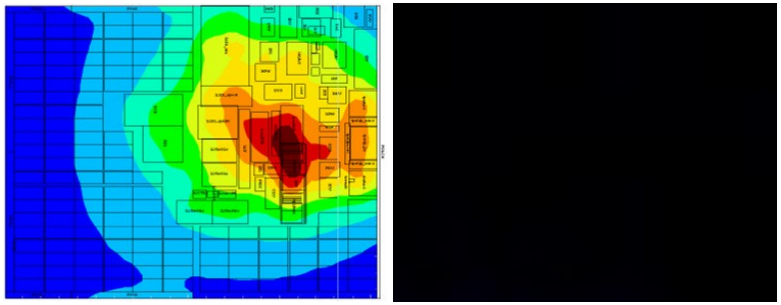
Challenges: production

- Increased variability = end of deterministic era
  - Random & systematic (lithography, voltage, thermal)
  - Variation 30% speed, 20X in leakage

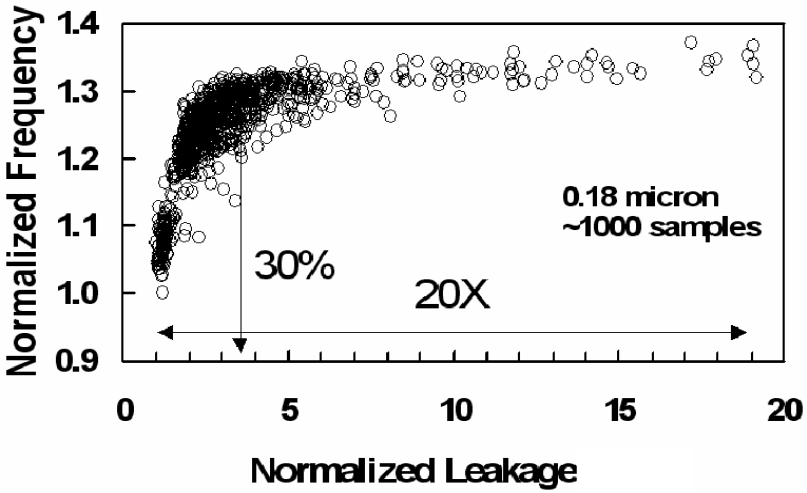


across wafers

between die



Temperature of a modern processor

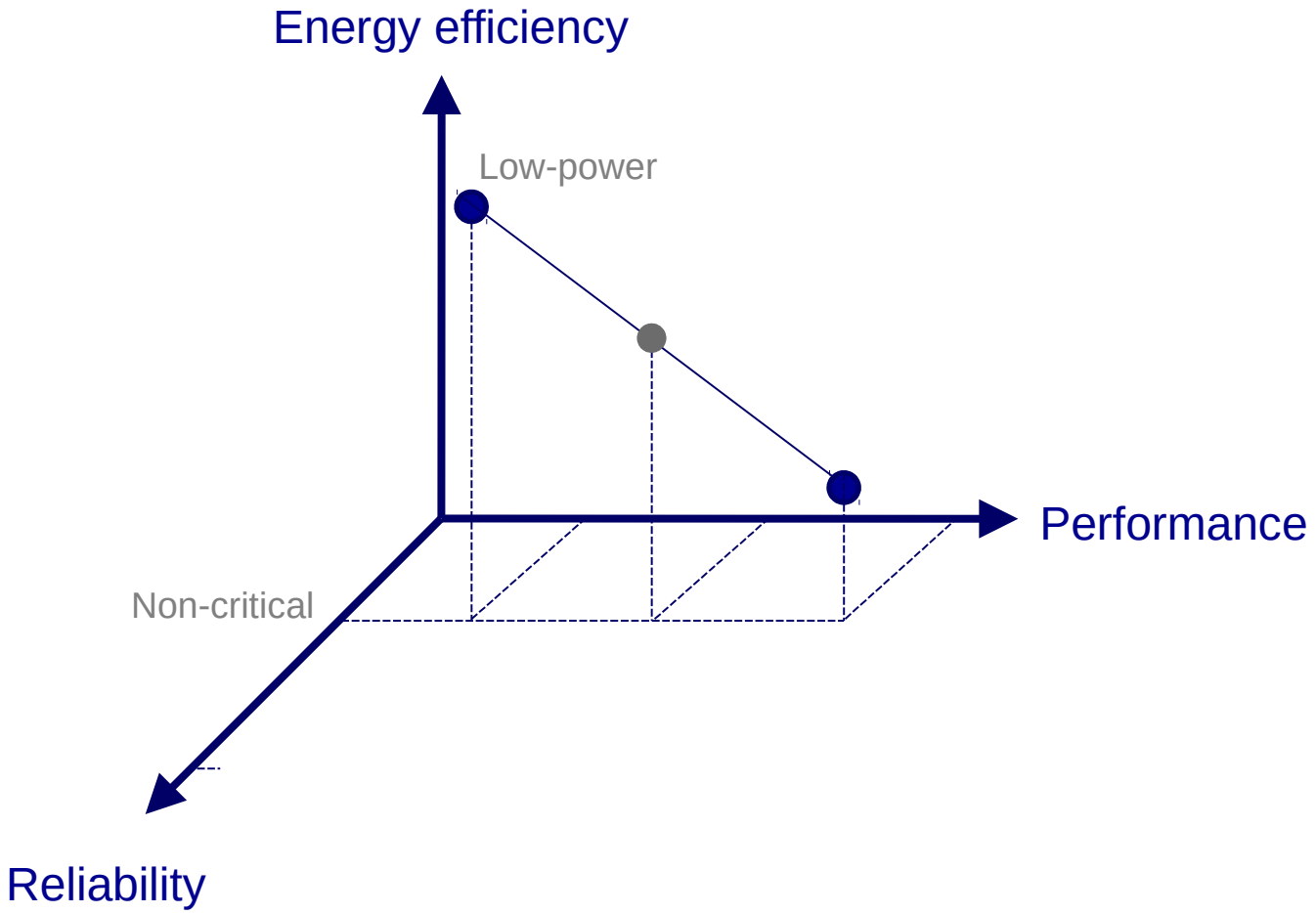


- Yield: design for manufacturability (DFM)
  - Higher defect densities and failure rates
  - Featured limited (style) vs. defect limited (area)

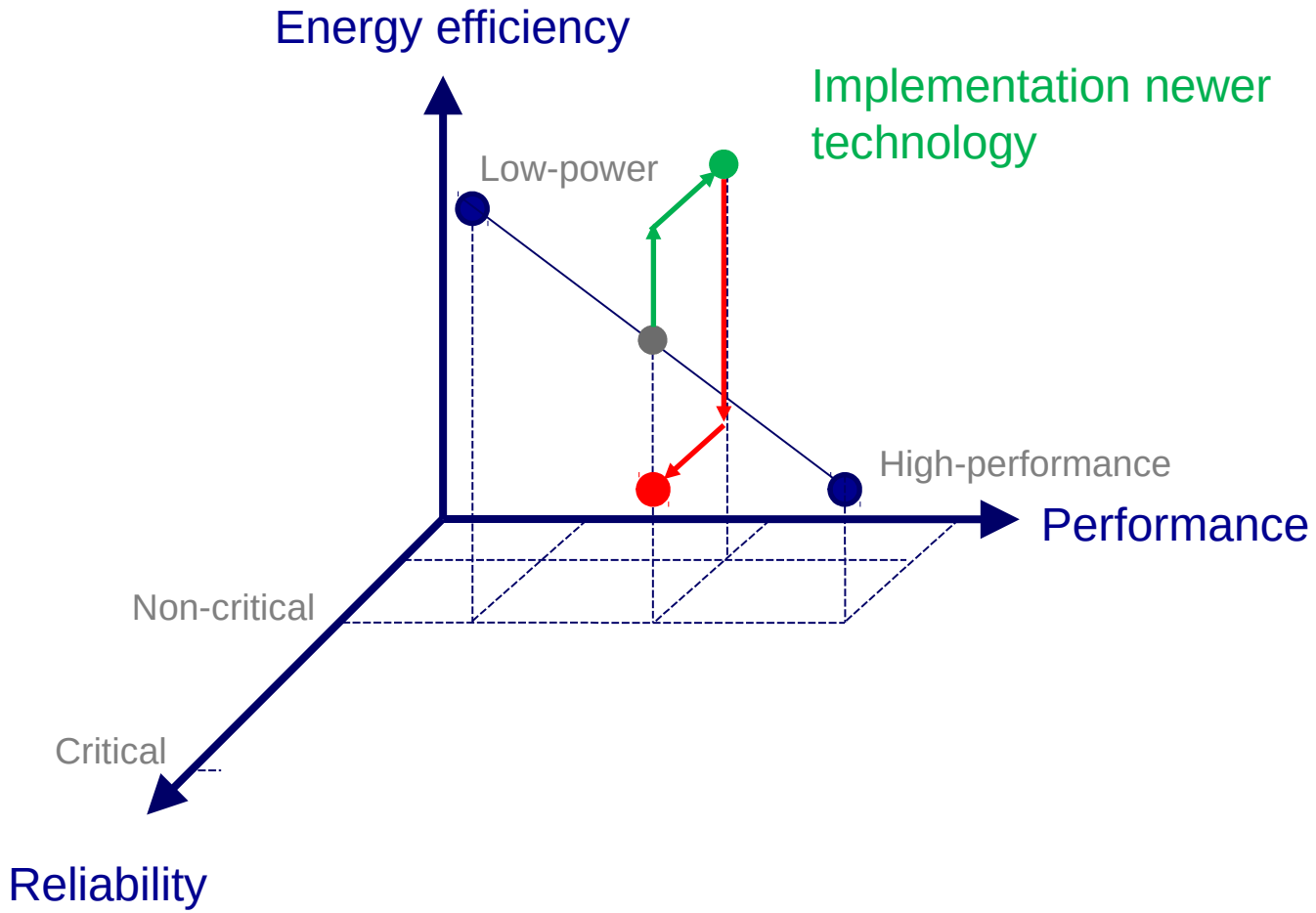
- Mask costs doubling each generation
  - $0.7M @90nm, $1.5M @65nm, $3.0M @45nm

Itanium II (1.5 GHz) (Madison processor)
J. Stinson, S.Rusu (Intel Corporation) http://videos.dac.com/videos/40th/41/41_3/41_3slides.pdf

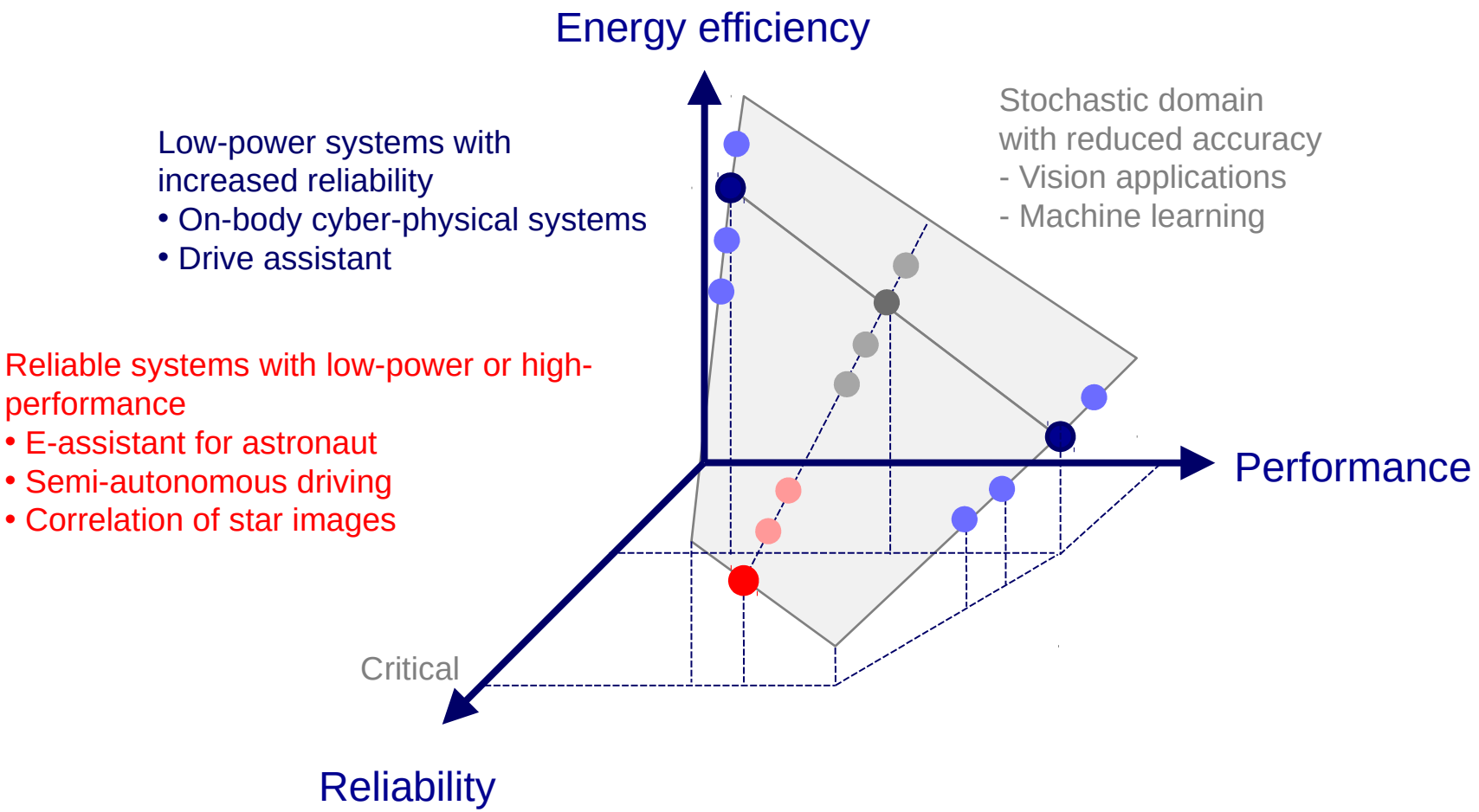
Trade-off Energy-Performance-Reliability



Trade-off Energy-Performance-Reliability



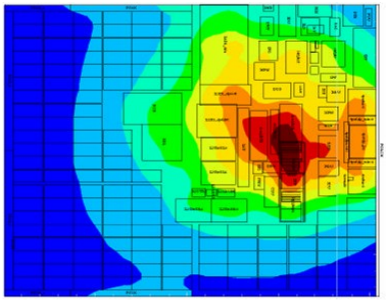
Trade-off Energy-Performance-Reliability



Taxonomy: error source

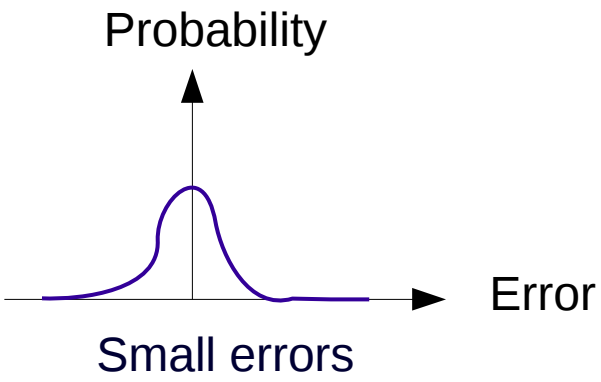
- **Approximate processing:** Approximate functionality at the functional level is introduced to simplify the hardware implementation
  - Can be modeled with Boolean logic

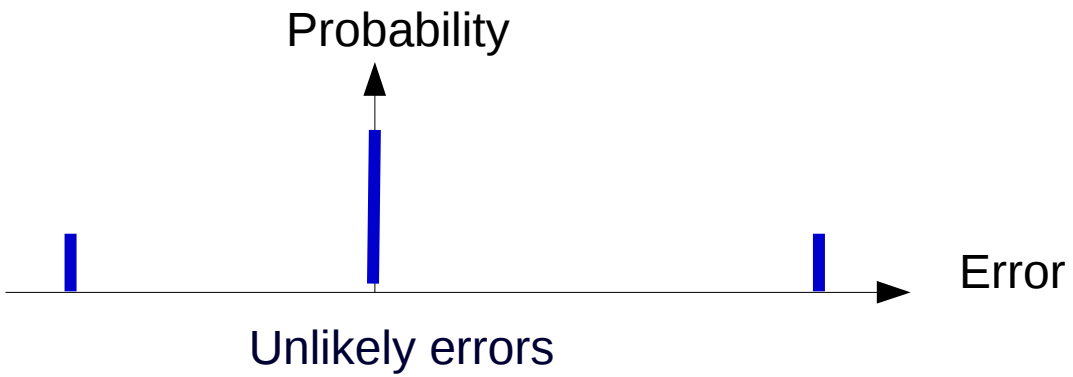
- **Stochastic processing:** Operation conditions (voltage, frequency,…) do not guaranty an error-free functionality. Additional hardware may be required to correct errors.
  - Needs modeling of physical effects



Taxonomy: error type

- **Small error:** The errors are small in magnitude although frequent. Errors are expected to be masked my processing noise.
- **Unlikely error:** Error are large but appear very small frequently, so that they are expected not to affect the application





Conclusion

Technology

- The reliability problem is getting more severe with newer technologies, where guaranteeing proper functionality is getting increasing expensive in terms of energy consumption.

Applications

- A new kind of applications requiring a softer trade-off between energy efficiency and robustness is emerging. Current approaches are insufficient to handle this trade-off efficiently..

Types

- The two conceptual alternatives are approximate processing and stochastic processing.

Approximate and stochastic computing: opportunities for processing hardware architectures

- Introduction and motivation
- Approximate processing
- Stochastic processing

Overview

- The idea of approximate the logic functionality of a unit is so general that can be equally applied to software and hardware.
- Especially well suited for data intensive applications:
  - DCT/IDCT transformation, motion estimation, CORDIC
- Can be applied at different levels of abstraction:
  - Hardware efficiency improvement in DSP units
  - Functional approximation, number representation
  - Prunning in CNNs, in communication decoders,...

Error metrics

- Error metrics quantify the quality of the approximation?

$$PE = E[\delta(\varepsilon)] = \sum_j \delta(\varepsilon_j) Pr[\varepsilon_j] \ ,$$

$$\sigma = \sqrt{E[(\varepsilon - \mu)^2]} = \sqrt{\sum_j (\varepsilon_j - \mu)^2 Pr[\varepsilon_j]}$$

$$MSE = E[\varepsilon^2] = \mu^2 + \sigma^2 \ ,$$

$$MAE = E[|\varepsilon|] = \sum_j |\varepsilon_j| Pr[\varepsilon_j] \ ,$$

Which one should you use? ….

Error metrics

- Classical error metrics are misleading

$$PE = E[\delta(\varepsilon)] = \sum_j \delta(\varepsilon_j) Pr[\varepsilon_j] \ ,$$

$$\sigma = \sqrt{E[(\varepsilon - \mu)^2]} = \sqrt{\sum_j (\varepsilon_j - \mu)^2 Pr[\varepsilon_j]}$$

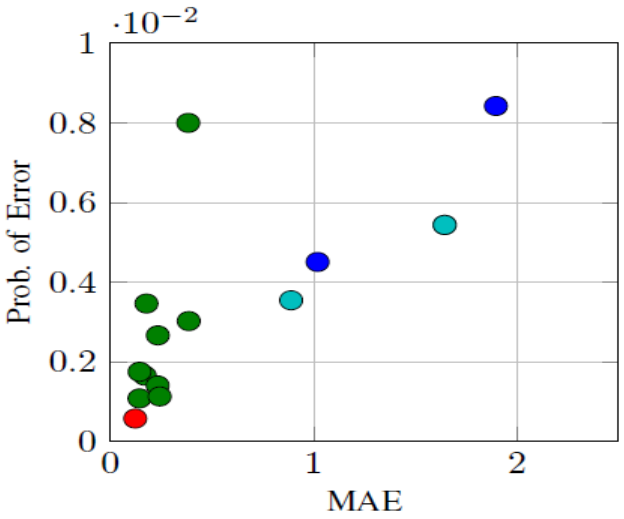
$$MSE = E[\varepsilon^2] = \mu^2 + \sigma^2 \ ,$$

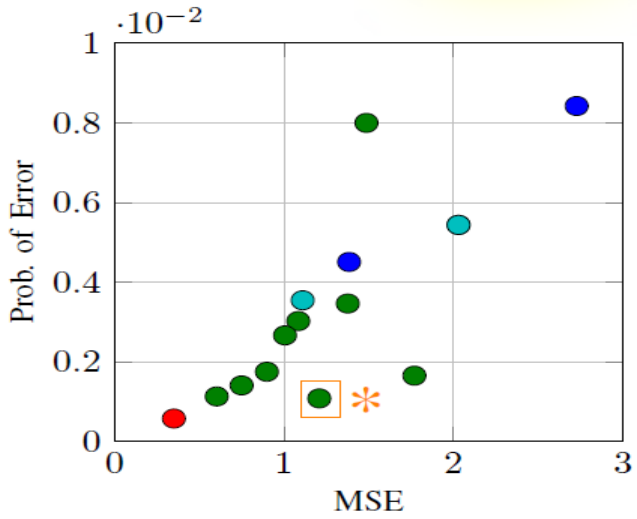
$$MAE = E[|\varepsilon|] = \sum_j |\varepsilon_j| Pr[\varepsilon_j] \ ,$$

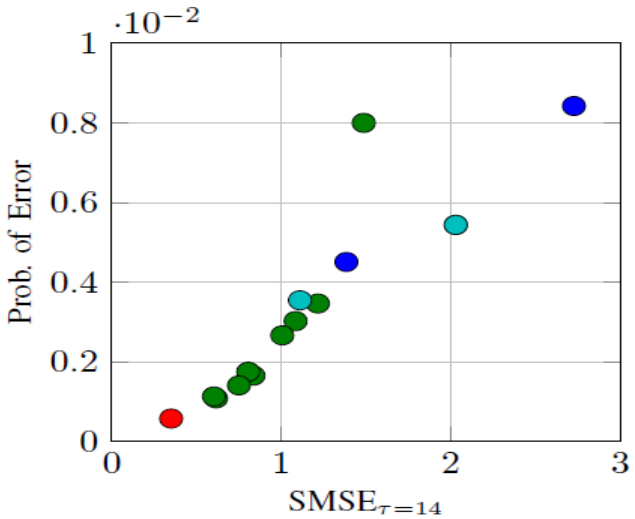
- Saturated metric more robust and more fidelity

$$SMSE_\tau = E[\min(\tau, |\varepsilon|)^2]$$

Final error in image application vs metric





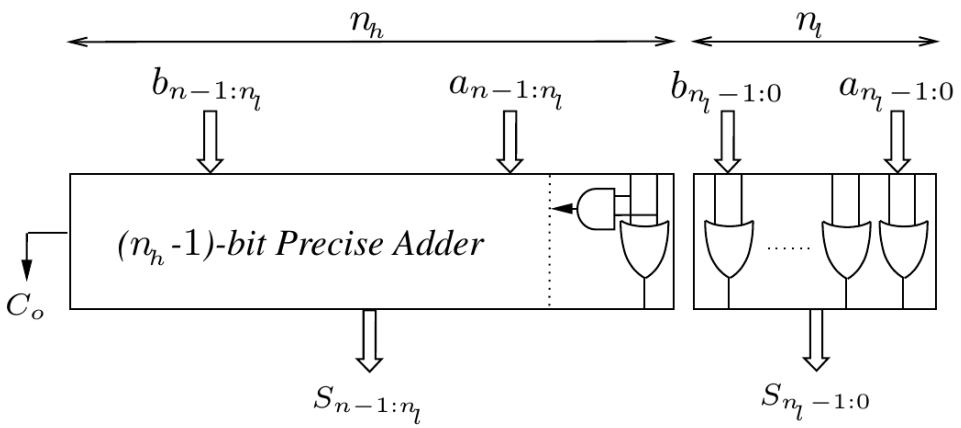


Examples of approximate adders

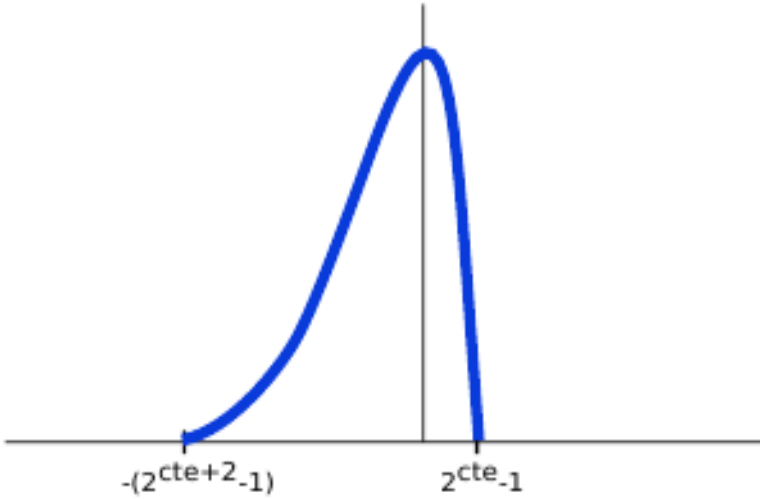
- Small error
  - Lower-part OR Adder (LOA)
  - Optimal Lower-part Constant-OR Adder (OLOCA)
- Uncommon errors
  - Equally segmented adder (ESA)
  - Generic Accuracy Configurable Adder (GeAr)
  - Error Tolerant Adder (ETA-II)
- Hybrid approaches

LOA

- Approximate FA in adder by OR gates



Architecture LOA



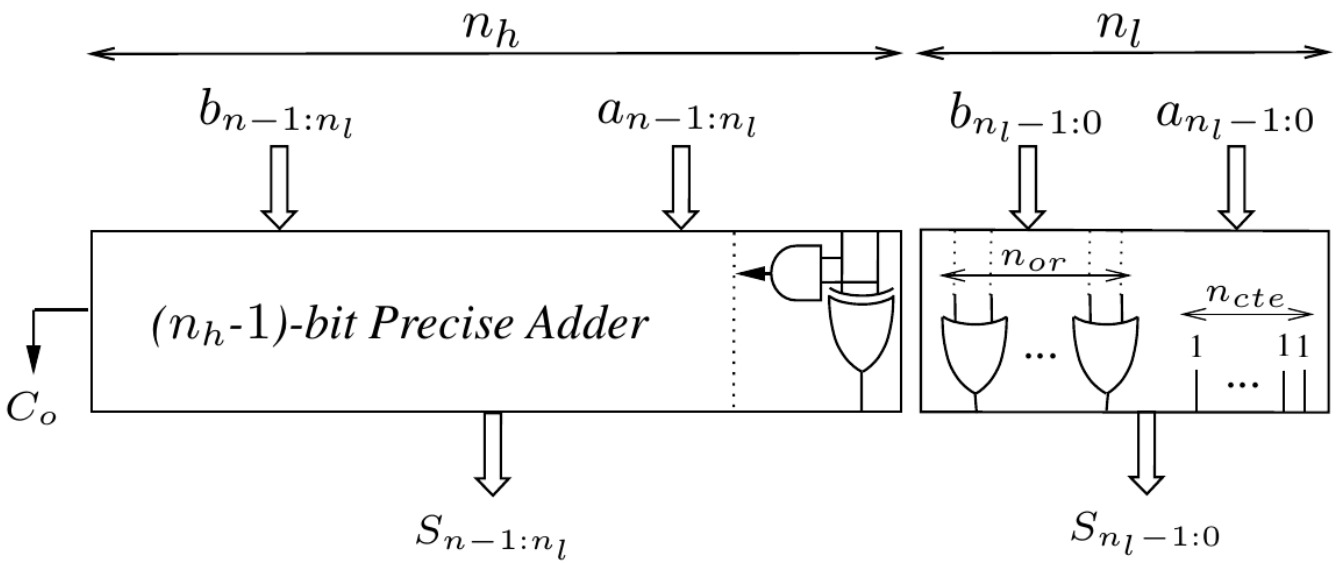
Error histogram LOA

OLOCA

- Optimize approximation

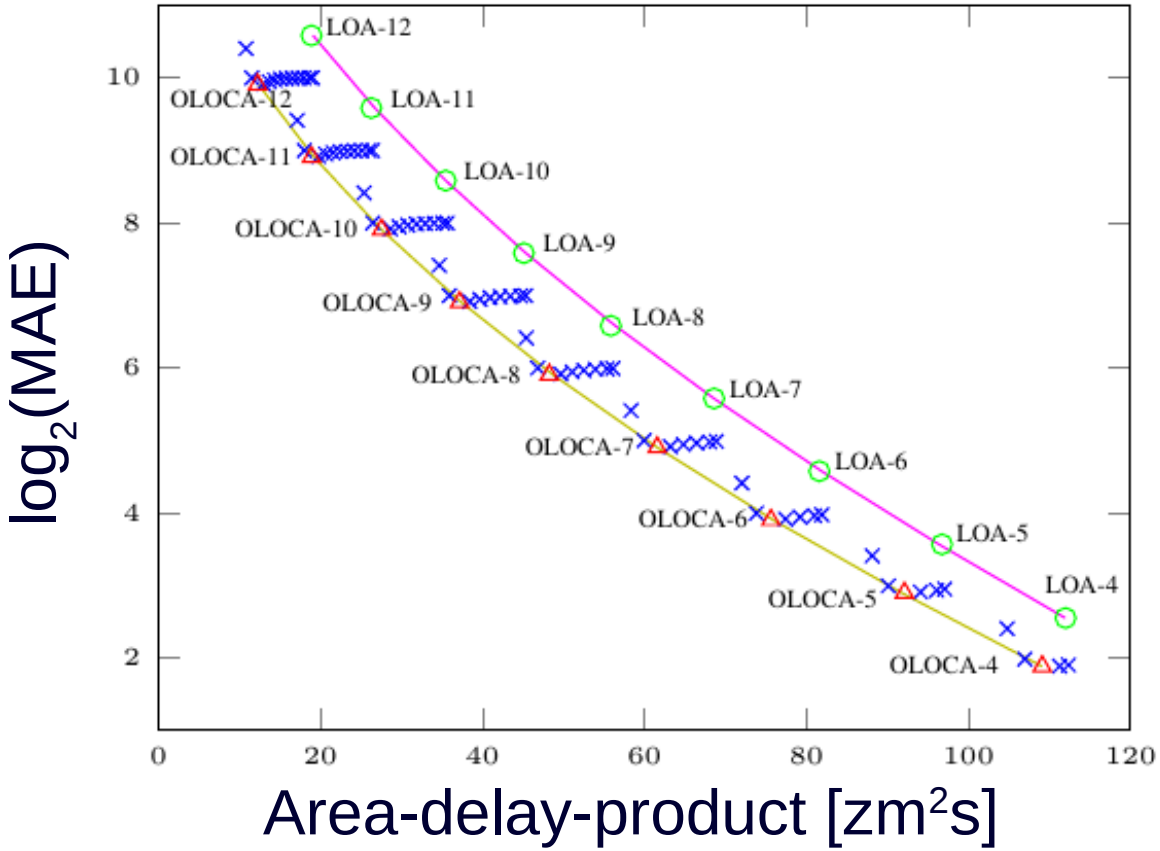
$$MSE_T = \sum_{i=0}^{n_l} \hat{\sigma}_i^2 2^{2i} + \left(\sum_{i=0}^{n_l} \hat{\mu}_i 2^i \right)^2$$

	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{MSE}$	$\hat{A}$	$\hat{D}$
AND	$-3/4$	$3/16$	$3/4$	1	1
OR	$-1/4$	$3/16$	$1/4$	1	1
Buffer	$-1/2$	$1/4$	$1/2$	0	0
Cte-0	-1	$1/2$	$3/2$	0	0
Cte-1	0	$1/2$	$1/2$	0	0



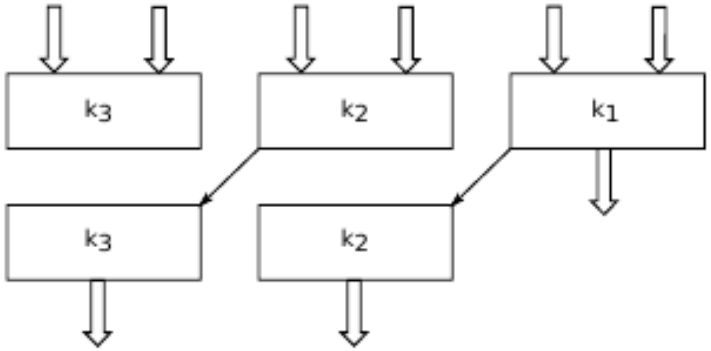
LOA vs. OLOCA

- OLOCA is optimal

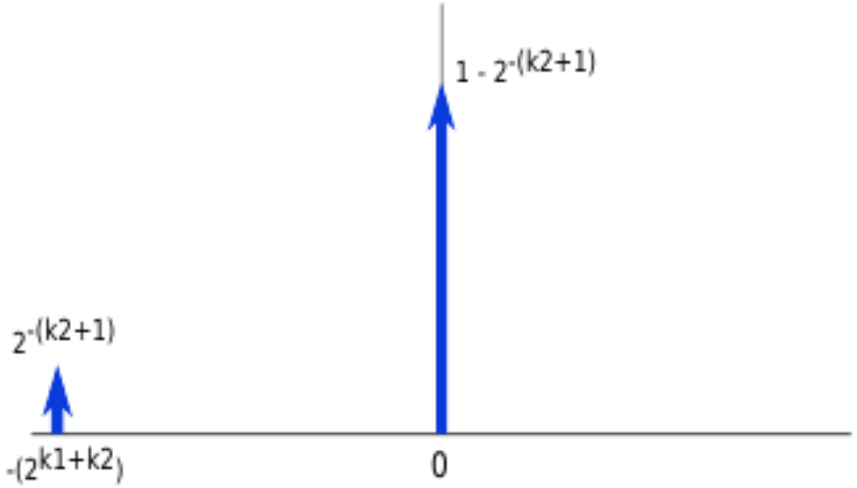


ESA and ETA-II

- ESA: Cut adder in segments to speed carry propagation
- ETA-II: Estimate carry using reduce set of bits



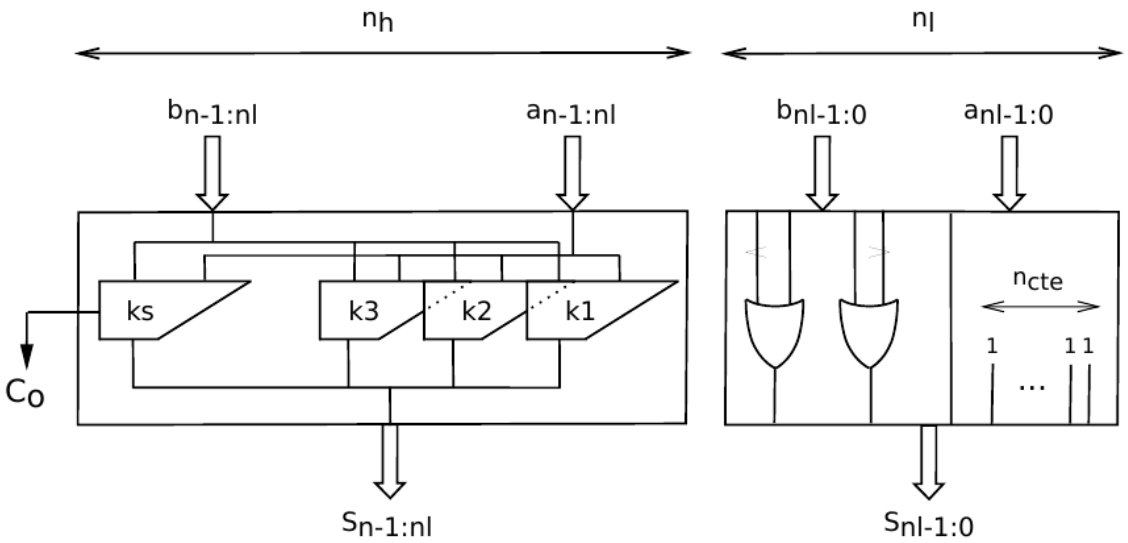
Architecture ETA-II



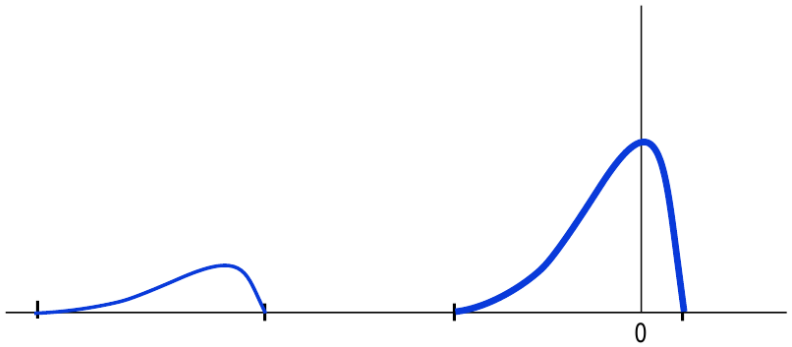
Error histogram ETA-II

Hybrid adder

- Combine the two error philosophies...



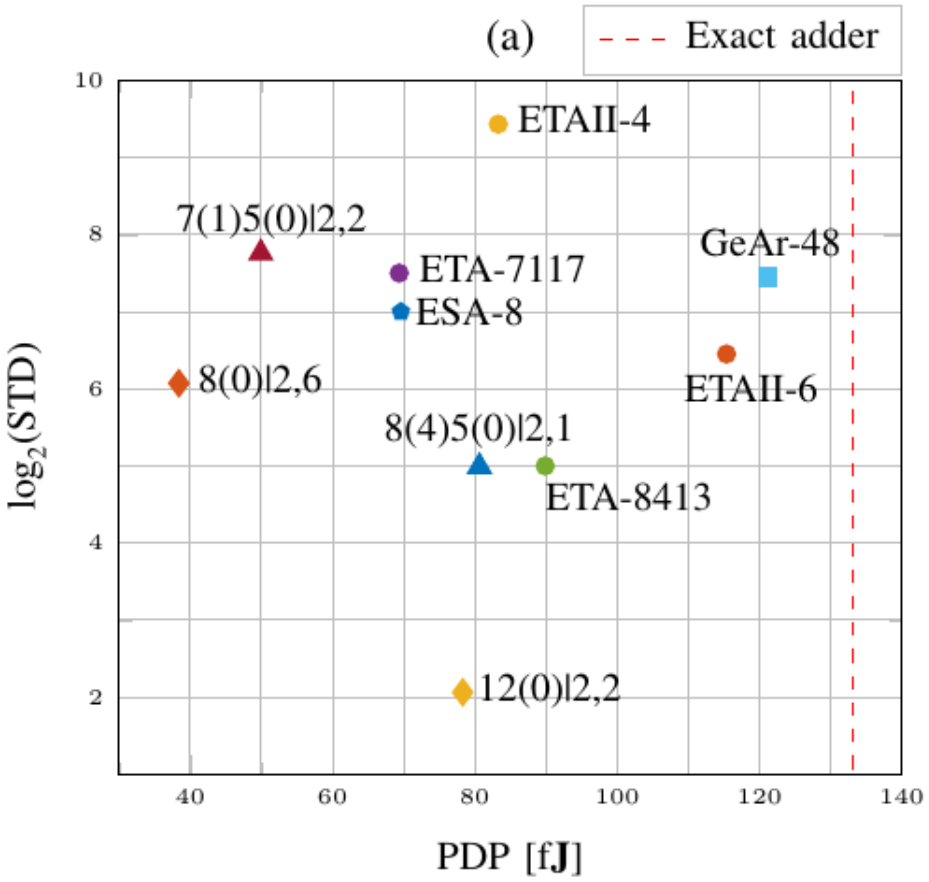
Architecture hybrid adder

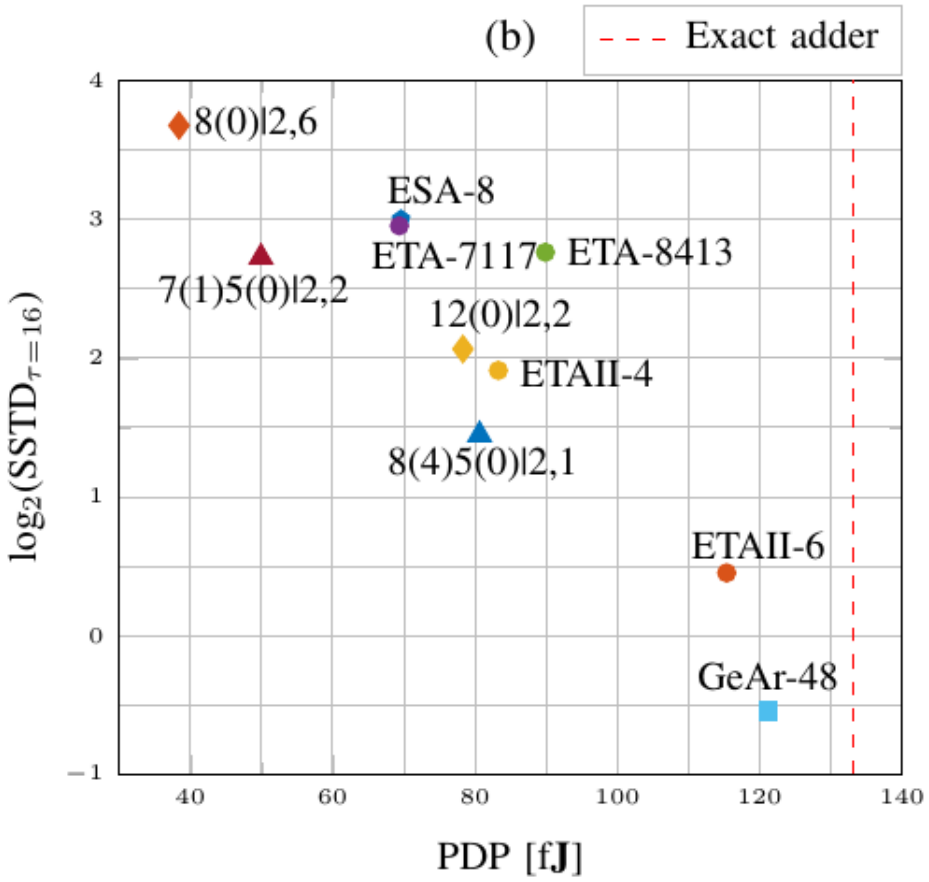


Error histogram hybrid adder

Comparison

- Non-trivial search of optimal… CAD tool required to find best architecture!!





Conclusions

Methodology

- Definition of quality metric for the approximate unit is fundamental.
- CAD tool can synthesize the optimal architecture

Architectures

- There are approaches with different error assumptions
- Best options: OLOCA (for small errors) and hybrid adders (in general)

Approximate and stochastic computing: opportunities for processing hardware architectures

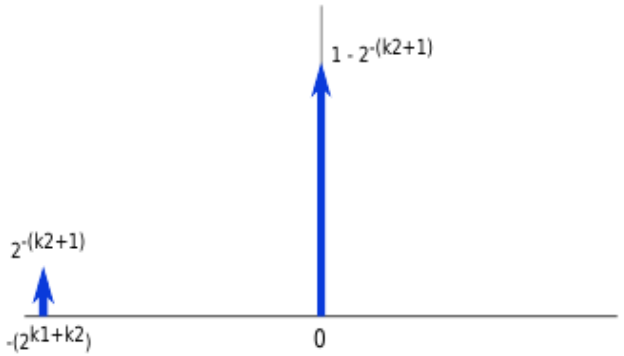
- Introduction and motivation
- Approximate processing
- Stochastic processing

Stochastic processing

- Rather than hiding hardware implementation constraints under expensive guardbands, designers can relax the traditional correctness constraints and deliberately expose hardware variability, obtaining significant processing performance improvements and energy benefits.
- Typical hardware constraints are those relatedto the maximum operating frequency (i.e., timing constraints) and power supply (i.e., operating voltage).

Errors in stochastic processing

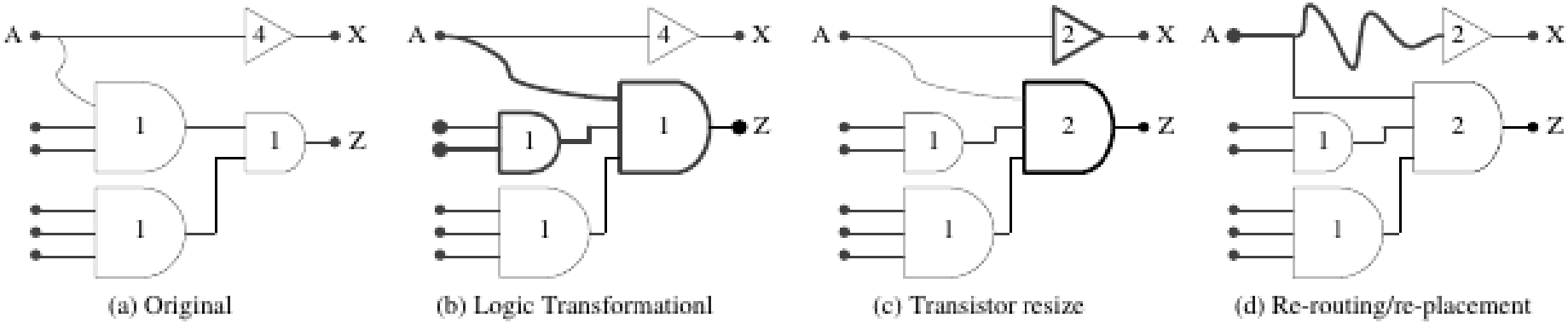
- Errors appear only when the critical path of the unit is active → very low probability
- When the errors appear, they can have a large magnitude



- Optimization
  - Reduce the probability of error
  - Reduce the impact of the error

Gate level

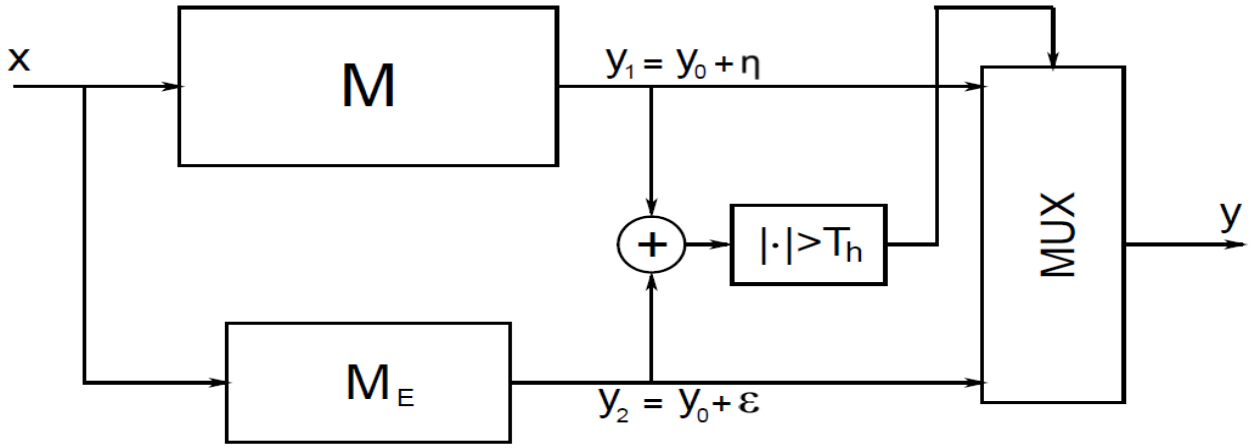
- Redistribute the timing slack during synthesis
  - Violated timing paths are identified during a gate-level simulation and are optimized iteratively until an error rate is below a specified threshold. Therefore, the main goal of these techniques is to
- Examples: Blue-shift, Dyna-tune...

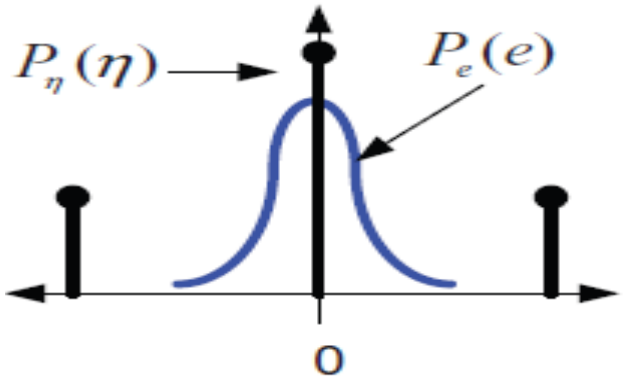


RTL level

- Error prediction
  - Telescopic units
- Error detection
  - Razor: Flip-flop with two sampling points
- Minimization of effect of errors with additional hardware
  - Example: Algorithmic Noise Tolerance ANT

ANT





- Error probability distributions of M and M_E different
- Estimator limits errors of main block

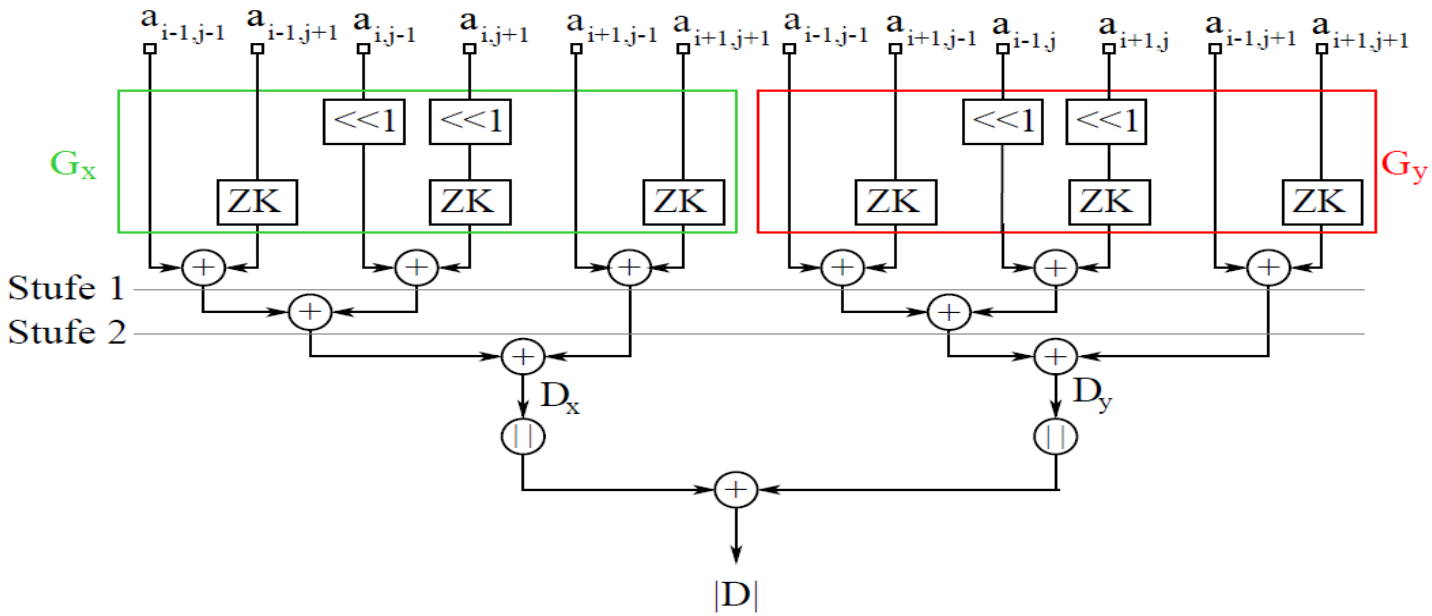
Application to image processing

- Soebel operator
  - Edge detection

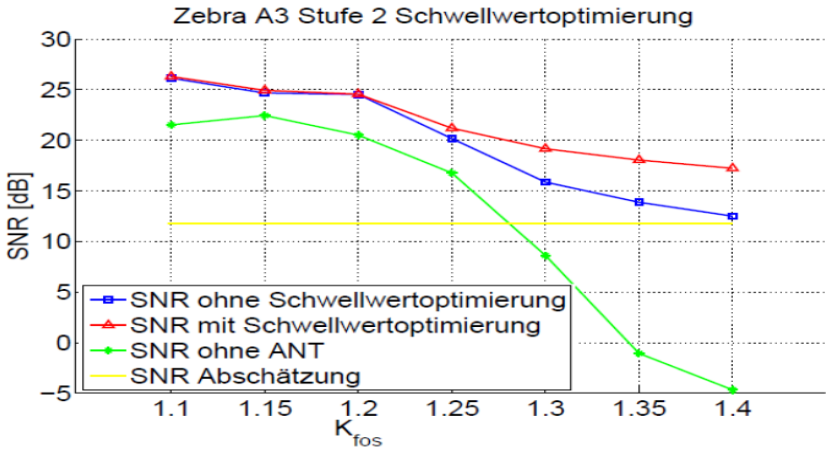
$$G_x = \begin{pmatrix} 1 & 0 & -1 \\ a & 0 & -a \\ 1 & 0 & -1 \end{pmatrix} \qquad G_y = \begin{pmatrix} 1 & a & 1 \\ 0 & 0 & 0 \\ -1 & -a & -1 \end{pmatrix}$$

$$a = 2$$

$$|D_{ij}| = \sqrt{D_{x,a_{ij}}^2 + D_{y,a_{ij}}^2} \approx |D_{x,a_{ij}}| + |D_{y,a_{ij}}|$$



Architecture with ANT

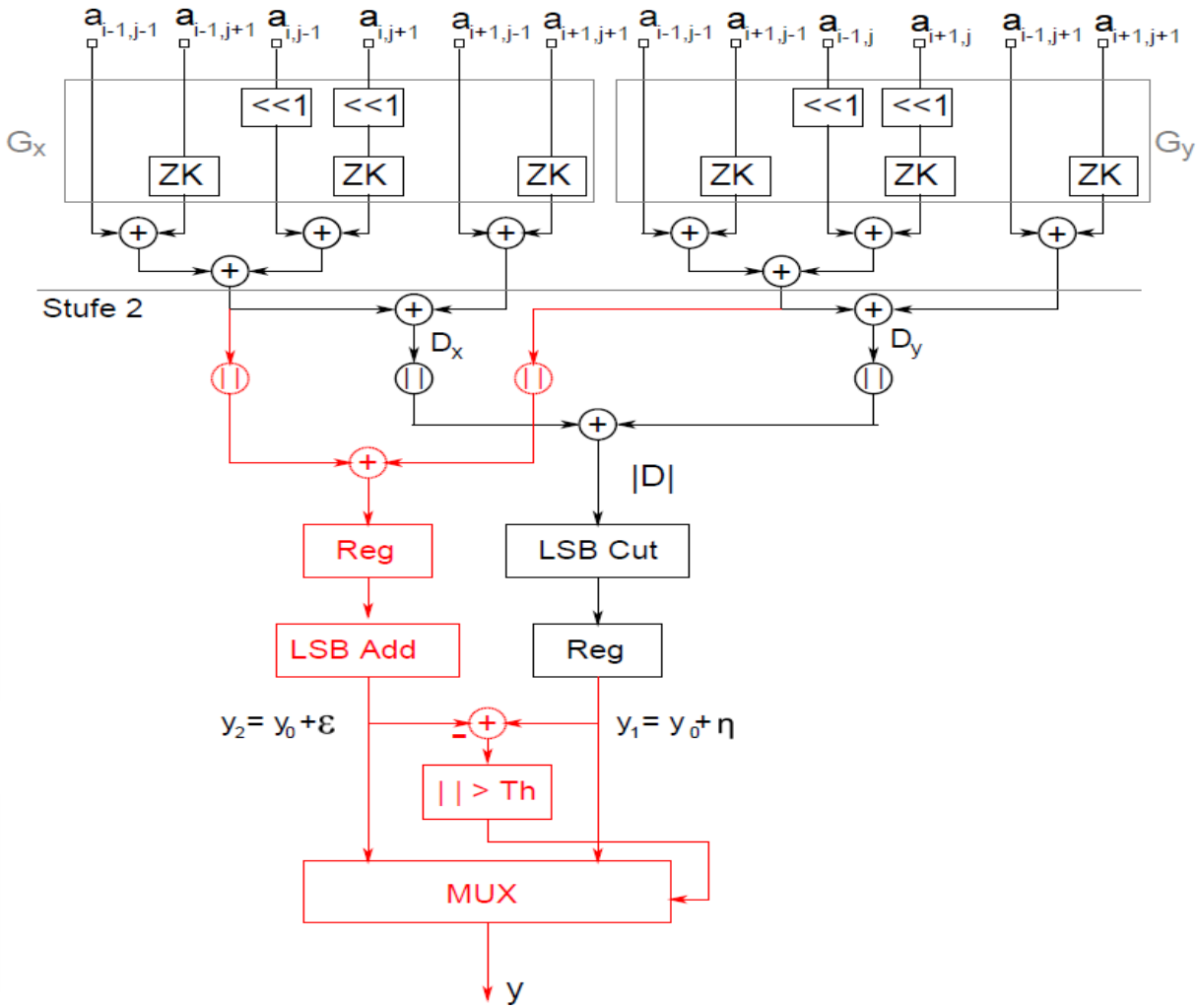


a) Originalbild

b) ohne ANT K_{fos}=1.0

c) ohne ANT K_{fos}=1.4

d) mit ANT K_{fos}=1.4

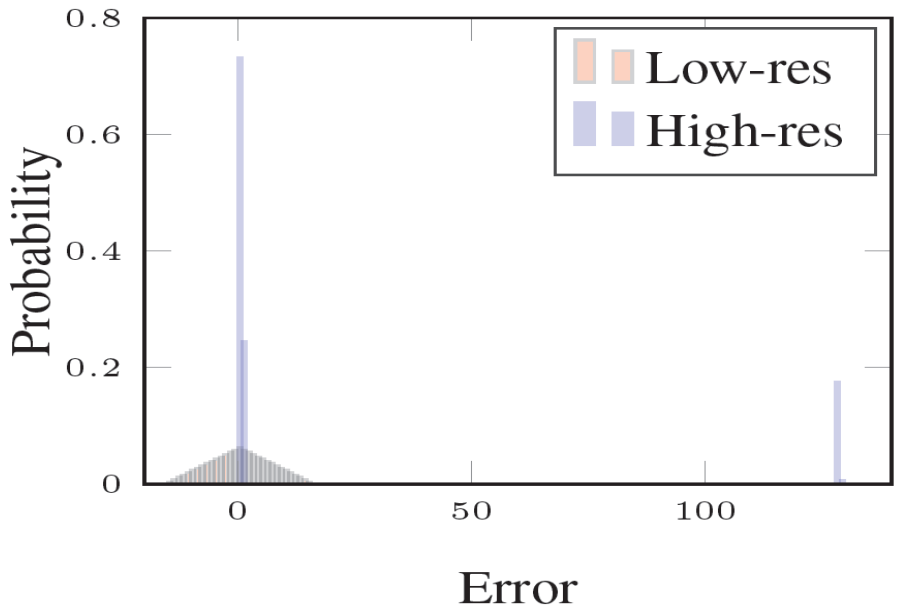


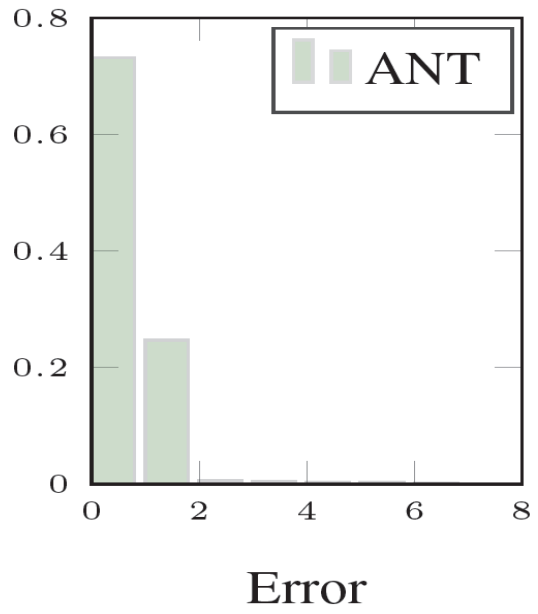
ANT with approximate adders

- Error of ANT using a high-resolution approximate adder and low-resolution as a replica

Adder	ANT input MSE	ANT input PE	ANT output MSE
ESA-4	10.9545	0.4688	7.4330
ETAII-3	14.9666	0.0547	0.8101
Hybrid	19.6405	0.2676	0.7539

Input and output error metrics in ANT





Input and output error histogram in ANT

Conclusions

Methodology

- Stochastic approaches need model physical layer and functionality
- Technique can be applied at different levels of abstraction

Architectures

- Understand the errors to reduce them!

Conclusions

Imprecise processing:

- Stochastic and approximate technique can reduce dramatically complexity of processing unit
- Make errors first-class citizens

Issues for adoption:

- Need for new methodologies
- Need for new architectures

Approximate and stochastic computing: opportunities for processing hardware architectures

Prof. Alberto Garcia-Ortiz

Tensilica Day, Hannover 2019

IEEE

MOCAST

INTERNATIONAL CONFERENCE ON
MODERN CIRCUITS AND SYSTEMS TECHNOLOGIES

HOME | TOPICS | PAPER SUBMISSION | REGISTRATION | CONTACT MOCAST

www.ids.uni-bremen.de/mocast2020

11 - 13 May 2020
Bremen, Germany