Design Techniques for Energy-Quality Scalable Digital Systems

Candidate: Daniele Jahier Pagliari
Supervisors: Enrico Macii, Massimo Poncino

Turin, 18 May 2018
Doctoral Program in Computer and Control Engineering (30th Cycle)
Outline

• Introduction and Motivation

• EQ Scalable Design Techniques for Processing Hardware

• EQ Scalable Design Techniques for Serial Interconnects

• EQ Scalable Design Techniques for OLED displays

• Conclusions and Future Work
• Energy efficiency is a key objective in modern digital systems

Battery operated devices
Energy harvesting

Mobile, IoT

Diminishing returns of classic techniques

Technology scaling
Voltage Scaling

• A lot of energy is spent in ensuring that the system performs **reliable, precise and accurate** operations (e.g. floating point, redundancy, etc.)
• Many modern computing applications are error tolerant (or resilient)

• For these applications, controlled errors in internal operations do not have a dramatic impact on final output quality

• Error tolerance can have different origins:
• **Noisy data** (e.g. from sensors) are affected by environmental noise
  • Errors can be tolerated as long as their effect on outputs is negligible w.r.t. the effect of noise

• **Redundant data** do not add information
  • Their computation can be approximated or skipped without degrading output quality
• The definition of **correct outputs** can be fuzzy or informal
  • If correct outputs are unknown (e.g. optimization problem)
  • If multiple outputs are equivalently good (e.g. Google search)

• Many applications have **human users**
  • Small or rare errors (in time and space) are not perceived by our sense organs
• Some **computational patterns** naturally reduce the effect of errors

• **Iterative refinement** steps converge to correct results even in presence of (controlled) errors.
  - E.g. Gradient Descent

• **Statistical aggregation** tends to reduce the effect of errors
  - E.g. Data mining, clustering, etc.
• Purposely introducing errors (i.e. relaxing the precision, reliability and accuracy of operations) can yield energy benefits:
  • Reducing data-path precision
  • Reducing design margins
  • Eliminating redundancy
  • Evaluating approximate functions
  • Etc.

• Energy-Quality (EQ) scalable design techniques exploit this tradeoff systematically for error tolerant applications

• The available “quality slack” depends on: task, context and inputs
• EQ Scalable System Architecture:
Motivation

• Our work focuses on three important aspects which are given little consideration in the EQ scalable design state-of-the-art:

1. Generality:
   • Holistic EQ scalability (not limited to processing)
   • Runtime quality-configurability

2. Automation and integration:
   • Compatibility with EDA tools
   • Compatibility with standard protocols

3. Focus on overheads:
   • Avoid implementations that offset energy gains
Outline

• Introduction and Motivation

• EQ Scalable Design Techniques for Processing Hardware

• EQ Scalable Design Techniques for Serial Interconnects

• EQ Scalable Design Techniques for OLED displays

• Conclusions and Future Work
**Target:** Hardware (HW) data-path modules

EQ Scalable Data-Path HW

- Reduced-Precision Redundancy
- Dynamic Voltage and Accuracy Scaling
  - Two Variants

**Objectives:**
- Automation (integration with EDA tools)
- Generality
• Reduced-Precision Redundancy:
  • Voltage Over-Scaling (VOS) on the original HW block (MDSP)
  • Error-Control (EC) block to mitigate the effect of timing errors

• EC block structure:
  • Estimator of the error-free output:
    • Implemented as a reduced-precision Replica of the MDSP
  • Decision block to select between MDSP and Replica outputs
EDA Flow for Reduced-Precision Redundancy

Limitations of classic RPR implementations:

1. Simplified and unrealistic assumptions on the input statistics
   - All timing path activations assumed equally probable.

1. No integration with standard EDA tools:
   - Simplified VOS timing degradation model
   - Ad-hoc replica implementation
EDA Flow for Reduced-Precision Redundancy

- **Goal of the proposed method:**
  - Automatically add RPR to the existing *gate-level netlist* of a data-path HW block
  - Under a user-defined *minimum output quality* constraint

- **Features:**
  1. Functionality *agnostic*
  2. Fully *automatic* and integrated with EDA tools
  3. Based on back-annotated *simulations* (with realistic models)
Experimental Results:

- Accurate consideration of input statistics is fundamental.

FIR filter RPR power savings for two input sets under identical conditions and constraints.

Power savings under realistic quality constraints:

The proposed method is general.

Area overheads under realistic quality constraints:

Design Techniques for Energy-Quality Scalable Digital Systems

December 10, 2018
Dynamic (Voltage) and Accuracy Scaling:

- Advantages:
  - Based on technological knobs only, no architectural modification
  - General
  - Low overheads
  - Many energy/quality configurations

- Limitations:
  - Integration with standard EDA flows
  - Slack does not increase as expected ("wall of slack")
  - Limited power benefits!

Design Techniques for Energy-Quality Scalable Digital Systems  
December 10, 2018
EDA Flows based on DAS/DVAS

- Solution 1: **Combination with fine-grain \( V_{th} \) tuning**
  - Split the HW block into \( V_{th} \) “domains”
  - Use \( V_{th} \) tuning to speed-up **timing-critical sections** of the HW for each precision
  - Implemented on FDSOI using **back-biasing**

Design Techniques for Energy-Quality Scalable Digital Systems
• **Solution 2: Application-driven Synthesis Flow**
  
  • Use **multi-scenario** optimization to prevent the wall of slack
  • Take into account the application-dependent **usage frequency** of each precision
EDA Flows based on DAS/DVAS

- Experimental Results (Solution 1):

  Design Techniques for Energy-Quality Scalable Digital Systems

  December 10, 2018
Qualitative comparison of the proposed methods:

**RPR based solution**
- Rare unpredictable errors
- Two “quality modes”
- \( \approx 100\% \) area overhead

**DVAS based solutions**
- Systematic “errors”
- Many “quality modes”
- \( \approx 10\% \) area overhead

---

1. Jahier Pagliari et al, “An automated design flow for approximate circuits based on reduced precision redundancy”, *ICCD2015*
2. Jahier Pagliari et al, “A methodology for the design of dynamic accuracy operators by runtime back bias”, *DATE2017*
• Introduction and Motivation

• EQ Scalable Design Techniques for Processing Hardware

• EQ Scalable Design Techniques for Serial Interconnects

• EQ Scalable Design Techniques for OLED displays

• Conclusions and Future Work
EQ Scalable Design of Serial Interconnects

• **Target:** Serial Buses
  • De facto standard for **sensors, actuators** and I/O controller interfaces
  • Higher frequencies, no jitter, reduced crosstalk, lower cost (# of pins)
  • **Power consumption:**
    • Mostly dynamic \( P_{chan} = \alpha C_{tot} V_{DD}^2 f \)
    • Energy reduction can be achieved by **reducing \( \alpha \rightarrow \text{data encoding!}**

• **Relevant?**
  • The serial transmission of a **12-bit datum** can **consume as much as** the execution of a **32-bit instruction!** (e.g. on a large off-chip PCB trace)
  • A system may include **tens of serial buses**
Error tolerant serial bus traces:

- Highly temporally correlated on average
- Often “bursty”: long almost constant (idle) sections and short (bursty) sections of fast and large variation

Proposed Encodings (ADE and Serial-T0): leverage the correlation and “burstiness” of data to introduce controlled approximations on encoded data with small impact on output quality.

Goals:
- Integration: compatibility with standard protocols (I2C, SPI, etc.)
- Overheads: encoding and decoding HW/SW do not offset the energy gains on the bus
Approximate Differential Encoding (ADE):

- Exploit the effectiveness of **Differential Encoding (DE)** for correlated data
- Combine it with quality-driven **LSB saturation** to improve savings

<table>
<thead>
<tr>
<th>Time</th>
<th>t-1</th>
<th>t</th>
<th>t+1</th>
<th>Input words</th>
<th>Codewords</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>101010</td>
<td>101011</td>
<td>101000</td>
<td>101010</td>
<td>00000000</td>
</tr>
</tbody>
</table>

14 Total Transitions

0 Total Transitions
• Serial-T0 (ST0):

• Exploit **idle sections** of bursty data for energy savings
• Selectively transmit the correct datum or a special **0-transitions pattern**
  (interpreted as “*repeat previous datum*”)

<table>
<thead>
<tr>
<th>Time</th>
<th>t-1</th>
<th>t</th>
<th>t+1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>010011</td>
<td>010101</td>
<td>101001</td>
</tr>
<tr>
<td>Input words</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>010011</td>
<td><strong>111111</strong></td>
<td>101001</td>
</tr>
<tr>
<td>Codewords</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Specific for ”bursty” signals (e.g. images)
• Experimental Results:

EQ tradeoff for accelerometer data

EQ tradeoff for RGB image data
Experimental Results:


EQ tradeoff for an OCR application
• Introduction and Motivation

• EQ Scalable Design Techniques for Processing Hardware

• EQ Scalable Design Techniques for Serial Interconnects

• EQ Scalable Design Techniques for OLED displays

• Conclusions and Future Work
EQ Scalable Image Transformations for OLEDs

- **Target**: OLED Displays
  - Composed of emissive devices
  - Image-dependent power consumption
- **New dimension**:
  - Trade-off power for image “error”

- **State-of-the-art Algorithms**:

  For each pixel \( Y \)
  \[ Y' = kY, \quad k < 1 \]

  **Brightness (luminance) scaling**

  Power saving + image (contrast) enhancement

  For each pixel \( Y \)
  \[ Y' = T(Y) \]

Design Techniques for Energy-Quality Scalable Digital Systems

December 10, 2018
EQ Scalable Image Transformations for OLEDs

• Limitations of the state-of-the-art:
  • High-complexity (nonlinear optimization, histogram processing)
  • No implementation overheads analysis

• Proposed Techniques:
  • Low-overhead Adaptive Brightness Scaling (LABS)
  • Low-overhead Adaptive Power Saving and contrast Enhancement (LAPSE)

• Goals:
  • Automation: plug-and-play frameworks based on regression models trained with representative images
  • Overheads: implementable in SW or HW, in real-time, with low energy consumption.
• **Low-overhead Adaptive Brightness Scaling (LABS):**
  • Adaptive brightness scaling: change the scaling factor $k$ depending on the image.

  \[ Y' = kY \]

  • Co-optimize **power saving** and **image alteration**: Power-Similarity Product (PSP)

  \[ k_{opt} \propto \frac{1}{\sum Y} \]
EQ Scalable Image Transformations for OLEDs

- **Low-overhead Adaptive Brightness Scaling (LABS):**
  - Linear regression to fit optimal scaling factor to image luminance
  - Trained offline with representative images
  - Online transformation becomes $O(#\text{pixels})$ and only involves simple operations
• LABS Experimental Results:

Average savings (MSSIM ≈ 0.93)
EQ Scalable Image Transformations for OLEDs

- Low-overhead Adaptive Power Saving and Contrast Enhancement (LAPSE):
  - Observation: state-of-the-art transformations can be approximated by a 3rd order polynomial of the pixels luminance

\[
T(Y) = a_3 Y^3 + a_2 Y^2 + a_1 Y
\]

Easy to implement in SW and HW

Goal: minimize power and maximize contrast under a maximum alteration (MSSIM) constraint
• **Low-overhead Adaptive Power Saving and Contrast Enhancement (LAPSE):**
  • Training-based approach similar to LABS
  • Different objective and constraints

**EQ Scalable Image Transformations for OLEDs**

**Offline phase**

**Online phase**
• LAPSE Experimental Results:

- **Input**: 
- **State-of-the-art technique (PCCE)**: Saving 61.6%, MSSIM 0.692
- **LAPSE**: Saving 59.5%, MSSIM 0.799

- **Input**: 
- **State-of-the-art technique (PCCE)**: Saving 60.3%, MSSIM 0.795
- **LAPSE**: Saving 67.4%, MSSIM 0.791

- **Input**: 
- **State-of-the-art technique (PCCE)**: Saving 60.79%, MSSIM 0.860
- **LAPSE**: Saving 55.1%, MSSIM 0.691

*Similar results despite much lower complexity*
LAPSE Experimental Results:

<table>
<thead>
<tr>
<th>Image Size</th>
<th>SW Ex. Time [ms]</th>
<th>HW Ex. Time [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>512x512</td>
<td>6.49</td>
<td>0.52</td>
</tr>
<tr>
<td>1280x1280</td>
<td>34.68</td>
<td>3.28</td>
</tr>
</tbody>
</table>


- Hardware energy overhead per image \(\approx 1000x\) smaller than OLED energy consumption
- (LABS implementation is even simpler!)
• Introduction and Motivation

• EQ Scalable Design Techniques for Processing Hardware

• EQ Scalable Design Techniques for Serial Interconnects

• EQ Scalable Design Techniques for OLED displays

• Conclusions and Future Work
Conclusions and Future Work

• A set of EQ scalable design techniques has been presented that:
  1. Target **different components** of a digital system, not limited to processing
  2. Consider in detail the **integration** with industrial best-practices (e.g. tools, standard protocols)
  3. When possible, favor **automated** and widely applicable (**general**) solutions
  4. Thoroughly evaluate and try to reduce the energy **overheads** associated with EQ scalability

• All proposed techniques allow runtime tuning of the energy-quality tradeoff:
  • Fundamental considering the **time-varying** nature of quality constraints
Conclusions and Future Work

• Future directions:

1. Apply the proposed techniques within complete applications (started):
   • E.g. use EQ scalable HW blocks for **machine learning** acceleration

2. Investigate **system-level** EQ scalable design:
   • Synergistic application of techniques for different components and abstraction levels
   • Implementing the EQ “**control loop**” becomes more complex