Chip-Scale Energy and Power... and Heat

Prof. Paul Hasler

Electrical and Computer Engineering Department, Georgia Tech University

The views and opinions presented by the invited speakers are their own and should not be interpreted as representing the official views of DARPA or DoD.

Approved For Public Release, Distribution Unlimited
Maximizing Computational Capability with Minimal Power

Georgia Institute of Technology, Electrical and Computer Engineering Department, Atlanta, GA, 30332


Approved for public release; distribution unlimited

Maximizing Computational Capability with Minimal Power

Paul Hasler
Professor, Georgia Institute of Technology

DARPA activity:
ISP, CT2WS, SyNAPSE, Healics, TEAM
Power Efficient Computing

Portable Devices
- battery powered
  (or less)
- larger systems
  minimize battery size / weight

Get as much computation as possible…

Cortical Neurons
- 1000’s of inputs,
- 1000’s of channel populations,
- one output

Equivalent computation ~ 400MMAC / neuron
(no learning / growth)

~ roughly 20pW / neuron

400MMAC / neuron at 20pW…
digital is quite far away (100mW)
analog VMM closer (100μW)
analog HMM / dendrites get close…

Custom Analog ~ 1000 – 10000
more efficient than Custom Digital
(Mead 1990)

Analog (VMM): 10MMAC/ μW
Digital: 4  MMAC / mW (DSP)

Useful Analog must be Programmable / Configurable

~ 200TMAC
< 500 neurons
~ 40kW (comp) with 2000 DSPs
Modern System Design

When building analog systems, we expect to build primitives at the basic algorithm level....

Analog = programmable and configurable.

How to get enough analog engineers

Hierarchy is a key ingredient to the success of the digital circuit, and, until recently, one reason why large analog designs have been difficult.

Approved For Public Release, Distribution Unlimited
Levels of Energy Efficiency

Subthreshold Transistor Operation

Programmable Circuits (FG transistors)
- Eliminate mismatch
- Programmability

Analog Signal Processing
- ~ x1000 improvement in power efficiency
- Wide accessibility

Moving analog approaches /conceptual framework to a system design approach, similar to digital’s system transformation in the 1970’s / 80’s.

- Large need for tools to compile / program these systems.
- Link most “useful” at system /sig processing level
- Education / training / foundational theory is critical for designing.

These techniques open further opportunities to utilize / explore biologically inspired techniques.
MOS Transistor Derivation

\[ I_{DS} = I_0 e^{\kappa V_g / U_T (\frac{V_3}{U_T} - \frac{V_D}{U_T})} \]

\[ = I_0 e^{\kappa (V_g - V_3) / U_T} \quad (V_{ds} > 4 U_T) \]

Subthreshold

Mismatch is significant: 10mV \( V_T \) shift
~ 50% bias current variation

\[ \kappa = 0.58680 \]
\[ I_0 = 1.2104 \text{fA} \]

As devices shrink, most of useful operating region is Subthreshold
Programmable Analog Transistors

- Standard CMOS
- Data retention: 
  $< 5 \mu V (0.5 \mu m) \ (10 \text{ year}, 300K)$
- Apps: Filters, Data converters, Regulators, etc.

Accuracy $\sim 0.1\%$ between 
$100pA - 1\mu A \sim 10e^{-}$

Write degradation (100$\mu$C):
- $V_{\text{tun}}$ increase less than 25%
- $V_{\text{inj}}$ negligible change
  ($100\mu$C is $>10^9$ complete FG rewrite)

Otherwise, need a DAC at every parameter and/or memory, etc.
Industrial Quality Programmable Analog ICs

Input Offset Voltage Reduced to ±25µV

V. Srinivasan, G. Serrano, J. Gray, and P. Hasler, CICC 2005, pp. 739-742. (Best paper CICC 2005)
Analog Signal Processing Techniques

Constant Q Filterbanks

Gaussian Mixture Models / VQ

Vector-Matrix Multiplication

Adaptive Filters
Computing in Memory

Memory

Input

Micro Processor
(Pipelined Multiplexed)

Computing Element

Input

Y = A * B

Computing Element

Memory

Approved For Public Release, Distribution Unlimited
Programmable Transform Imager

- Digital Output
  - Coefficient Generation
  - ADCs
- Readout Offset
  - VMM
- Block Selection
- Computational Pixel Element Array
  - Readout Circuitry
  - Block Selection

### Technology
- 0.35μm CMOS

### Array size
- 256 x 256

### Pixel size
- 6 μm x 6 μm

### Fill factor
- 38%

### Die area
- 4.5mm x 4.5mm

- Image Read
- Log compress
- DCT -> IDCT
- Original
- Measured
- Edge Enhance
- DCT Matrix
- 2D DCT
- Reconstruct

- Log compress
- DCT -> IDCT

- Image Read
Analog--Digital Signal Processing

CADSP = Cooperative Analog—Digital Signal Processing

Custom Analog ~ 1000 - 10000 more efficient than Custom Digital (Mead 1990)

- Analog (VMM): 10MMAC/ \mu W ($= 10\text{TMAC} / \text{W}$)
- Digital: 4 MMAC / mW (DSP)

**Digital and Analog SP Efficiency**

<table>
<thead>
<tr>
<th>Computation</th>
<th>MMAC/\mu W</th>
<th>Ratio to digital</th>
</tr>
</thead>
<tbody>
<tr>
<td>LowPowerDSPs</td>
<td>0.02 to 0.002</td>
<td>1</td>
</tr>
<tr>
<td>Analog VMM</td>
<td>1 to 30</td>
<td>1000</td>
</tr>
<tr>
<td>Analog Filterbanks</td>
<td>30 to 1000</td>
<td>10000</td>
</tr>
<tr>
<td>Analog VQ</td>
<td>1 to 10</td>
<td>300</td>
</tr>
<tr>
<td>Analog HMM</td>
<td>&gt;1000</td>
<td>&gt; 100000</td>
</tr>
</tbody>
</table>

- Analog SP Power
- 1000 to 10,000 improvement
- 20 year leap
- Gene's Law
Resolution for Analog / Digital Tradeoffs

![Graph showing the tradeoff between signal-to-noise ratio and cost for analog and digital systems. The graph plots log("Cost") against Signal-to-Noise (Bits of Resolution). The digital and analog cost curves are shown, with digital costs increasing more steeply.]

Lower digital cost
Lower analog cost

Analog filter bank (~FFT)

ADC (16bit) → FFT → Remaining DSP

~10bit SNR

DSP Application

[Vittoz95, Sarpeskar98]

[Kucic, et. al. 2001]
Reconfigurable Signal Processing

FPGAs – Large Configurability

Power: Just MAC engine
around 2-10MMAC/mW
Baseline static power ~ 0.5W to 1 W
Signal routing power / memory: ?

DSPs – Low Power Processing
- cell phones
  (processing < 30mW average)
- hearing aids (1 mW levels)
  (AMI / DSP factory)

Power: 54C series – 4MMAC/mW

Power does not include comm off chip
(i.e. accessing memory)

Power = \( \frac{1}{2} C Vdd^2 f \) for CMOS
Chip to Chip (10pF load min, 2.5V):
32uW/Mbit (dynamic)

Obtaining data for 4MMAC computation ~ 4mW

Innovation and Process Scaling moves solutions towards programmability and reconfigurability
Moving towards Configurable Analog

Useful Analog must be Programmable / Configurable

FPAA =
Field Programmable
Analog Arrays

Can be a prototyping tool,
early devices, or
final application

• RASP 1.x (2002)
  (T. Hall, P. Hasler, et. al, FPL, Sept. 2002.)

• RASP 2.x:
  RASP 2.5, 2.7: 2004-2007
    (C. Twigg & P. Hasler, CICC, 2006)
  - >50,000 Prog. Analog Devices
  - Used by > 100 Eng
  RASP 2.8x: 2008-
    (A. Basu, et. al, CICC, 2008)
  - Used by > 50 Eng and growing
  RASP 2.9x: 2009-

Custom versus FPGAs: x2-3 speed, x10 area, x100 power
Custom versus FPAAs: < x2 speed, < x2 area, < x2 power
Next Questions on FPAA

FAQ on Large-Scale FPAA

- Design time similar time for FPAA targeted and custom ICs
- Size can be similar to custom (programmable caps / I)
- Noise levels are similar to custom design
- Similar speed as custom upto routing fabric speed (~10-20MHz in 0.35um CMOS)
- Power levels often similar to custom solutions
- Techniques scale (~ ideal CMOS rules) with process shrink

<table>
<thead>
<tr>
<th>Node (nm)</th>
<th>Prog #s (M)</th>
<th>TMACs</th>
</tr>
</thead>
<tbody>
<tr>
<td>350</td>
<td>4.0</td>
<td>1</td>
</tr>
<tr>
<td>90</td>
<td>64.0</td>
<td>64</td>
</tr>
<tr>
<td>45</td>
<td>256.0</td>
<td>512</td>
</tr>
</tbody>
</table>

- Neuromorfix to commercialize FPAA technology

Compiled circuits include:
- n-th order filters / filterbanks, Capacitive summation / differencing,
- Ramp ADC, Algorithmic and Sigma-Delta ADCs, MP3 encoder, WTA,
- Analog Distributed Arithmatic, HMM classifiers, Van-der-pol Oscillator
Rapid Prototyping using FPAAs

RASP 2.7 PhotoReceptor Response

Paper Strip
FPAA Workshops (RASP 2.8x)

LA Workshop
USC Campus, May 2-7, 2008

>30 Participants.

CO Workshop
Telluride, July 2008

>20 Participants.

ATL Workshop, Oct 2008
>25 Participants.

Other workshops being planned:
Boston, SF, Orlando, DC?

GT Neuramorphic Classs
(Fall2008, >20 students)

Education / training / foundational theory is critical for designing.
Simulink FPAA Tool

FPAA library
- Winner Take All
- Vector Matrix Multiply
- Low Pass Filter
- High Pass Filter
- Voltage Reference
- Current Reference
- Sum
- Difference

Simulink
- Block Model
  - Parser
    - Library
      - MATLAB Struct
        - Sub Circuit
          - Nelist Generator
            - SPICE
              - Targeting Code
                - FPAA

Petre, et. al, ISCAS 2008]
Getting higher power efficiency: Neuromorphic Engineering

400MMAC / neuron at 20pW vs. digital (100mW)
and analog SP (100μW)

- Neuromorphic processing = event-based processing
  uses power only when useful signals are present
  (“always on” in sensors or further processing)

Programmability and Configurability empowers
neuromorphic design towards useful applications
in a reasonable timeframe.
- Address Event Representation (AER) / FPGAs
- FPAAs / FG devices –
  ~ sizes of largest custom neuro ICs

Can model pyramidal cells in configurable fabric in ~1mm² area with
realistic channel, dendrite, and synapse elements (power in nW level and decreasing)
Levels of Energy Efficiency

Subthreshold Transistor Operation

Programmable Circuits (FG transistors)
- Eliminate mismatch
- Programmability

Analog Signal Processing
- ~ x1000 improvement in power efficiency

Configurable Signal Processing
- Wide accessibility

Moving analog approaches /conceptual framework to a system design approach, similar to digital’s system transformation in the 1970’s / 80’s.
- Large need for tools to compile / program these systems.
- Link most “useful” at system /sig processing level
- Education / training / foundational theory is critical for designing.

These techniques open further opportunities to utilize / explore biologically inspired techniques

DARPA activity:
ISP, CT2WS, SyNAPSE, Healics, TEAM