THE POWER WALL –
ARE WE SCALING IT OR IS IT JUST GETTING HIGHER?

Scaling the Power Wall–
Revisiting the Low-Power Design Rule

March 2008

Jan M. Rabaey
Director Gigascale Systems Research Center (GSRC)
Research Center Berkeley Wireless (BWRC)
California, Berkeley

Jan M. Rabaey
Donald O. Pederson Distinguished Prof.
University of California at Berkeley

CADENCE LOW POWER SUMMIT, OCT 18 2012
The Emerging Information-Technology Platform

Infrastructural core

Mobile Access & Relay

The Swarm

The Cloud

[J. Rabaey, ASPDAC’08]
Information Processing –
It Is All About Energy ...!

Further progress in all aspects of future information technology platform requires continuing increase in energy efficiency!

The Compute Cloud

Mobiles

ENERGY-INTENSIVE

ENERGY-FRUGAL

Sensory Swarm

ENERGY-BOUNDRED

Further progress in all aspects of future information technology platform requires continuing increase in energy efficiency!
Almost Three Decades of Low-Power Design

- Eliminating waste
- Reducing energy by voltage scaling
- Architectural innovation
- Power/voltage management

The mantra’s:
slow, simple, many, dedicated, adaptive
But … We are Running out of Options

Obvious waste has been largely eliminated (…)

Technology scaling may not help much anymore

Mobile processors already use most known tricks …

Minimum energy point set by leakage

Quadcore Snapdragon (©Qualcomm)

Process variations and random upsets dictate noise and timing margins
Trends Flat-Lining

- Transistors (100000's)
- Power (W)
- Performance (GOPS)
- Efficiency (GOPS/W)

Limits on heat extraction
Stagnates performance growth
Lost voltage scaling → energy efficiency plateaus

Era of High Performance Computing  
Era of Energy-Efficient Computing  
(Courtesy: D. Sylvester, U. Mich)
How to Move Forward (or Lower)?

ISSCC® 2011
Electronics for Healthy Living
San Francisco Marriott Marquis | San Francisco, California, USA

Plenary Talks (Mon, Feb 21)
- Game-Changing Opportunities for Wireless Personal Healthcare and Lifestyle
  - Jo De Boeck
  - IMEC
- Eco-Friendly Semiconductor Technologies for Healthy Living
  - Oh-Hyun Kwon
  - Samsung Electronics
- New Interfaces to the Body Through Implantable System Integration
  - Steve Oesterle
  - Medtronic

Plenary Roundtable (Mon, Feb 21)
Beyond the Horizon: The Next 10x Reduction in Power – Challenges and Solutions
- Jan Rabaey
  - Professor
  - UC Berkeley
- Hugo De Man
  - Professor Emeritus
  - KU Leuven
- Mark Horowitz
  - Professor
  - Stanford University
- Takayasu Sakurai
  - Professor
  - University of Tokyo
- Asad Abidi
  - Professor
  - UCLA
- Dan Dobberpuhl
  - Consultant
- Hermann Eul
  - Executive VP
  - Infineon
- Kiyoo Itoh
  - Fellow
  - Hitachi
- Philippe Magarshack
  - Group VP
  - STMicroelectronics
- Jack Sun
  - CTO/VP R&D
  - TSMC
ISSCC 2011 Plenary Panel
The next 10x Reduction in Power
(J. Sun, D. Dopperpuhl, K. Itoh, A. Abidi, H. Eul, P. Magarshack)

EE Times
News & Analysis

Highlighting the magnitude of the power efficiency challenge, asked by Rabaey near the end of the panel discussion to identify the single most important factor in chip power efficiency in the years ahead, all six panelists selected a different one.

Suggestions included, among other things – better materials, transistors, architectures, algorithms, less conservative, energy-consuming guardbanding, and most compellingly, STMicro’s new “sense & react” vision where a design enjoys far needs of an ever-transient environment.” Undoubtedly, those attending DATE 2011 next month in
Some Observations

- All these things will surely play a roll
  Some of them are already in play …
- Don’t expect too much from technology or circuits in the next decade
  - In lieu of better transistors, sub- or near-threshold circuits will play a bigger role
  - Self-timed circuits may finally take a front row seat
- Nothing truly exciting in the architecture arena (in the short term)
  - More of the same (slow, concurrent, adaptive, dedicated)
  - But … cool ideas in the making
Jan’s Energy-Efficiency Roadmap

- Continued Voltage Scaling
- Energy-Proportional Systems
- Always-Optimal Circuits
- Aggressive Deployment
- Beyond Turing
Lowering Supply Voltage Only Option
(recoup performance through parallelism)

BUT: CMOS Has Minimum Energy Point Set by Leakage

![Graph showing energy vs. supply voltage]

- Energy (norm.)
- $V_{DD}$ (V)

- Total
- Switching
- Leakage

0.3V
Sub-Threshold Operation Leads to Minimum Energy/Operation

Sub-Threshold Operation Leads to Minimum Energy/Operation

Energy-Aware FFT Processor
[Chang, Chandrakasan, 2004]

Perpetual sensor system with solar harvesting
ARM Cortex M3
28 pJ/cycle @ 400 mV & 72 kHz
[Blaauw, ISSCC 2010]

But ... Minimum Energy Comes At Huge Performance Penalty and Increased Variability
Option 1: Back off a bit …

Near-threshold design

**Benefits of Near Threshold Voltage Operation**
Peak energy efficiencies at NTV and fine-grain power management

- **Potential For...**
  - More always-on / instant wake devices
  - Intelligent everyday devices with battery/solar powered CPUs
  - Longer battery lives for mobile computing
  - Scalable many-core chips for the datacenter
  - Meeting extreme-scale compute challenges

[H. Kaul et al, ISSCC08] [Ratner, IDF2011]

Challenge: Modeling!

Intel Claremont Near-Threshold Voltage IA Core (400-500 mV)
Option 2: Self-adaptation

- Reference design from UCLA
- \(2\mu W/\text{channel}\) in 90nm @ 0.55V

*V. Karkare, ASSCC09

Minimize the leakage energy through asynchronous self-timed operation

Reduce logic gate leakage by exploiting the stack effect

Realize robust computations at a low supply voltage

A 0.05\(\mu\)m\(^2\), 0.25\(\mu\)W/\text{channel} self-timed spike-sorting DSP operating at 0.25V [VLSI12]
Emerging devices may offer some relief by lowering minimum energy point

Improved Sub-threshold Slope

NEMS Relay Logic
[King, Alon, ISSCC10]

Nanowire FET

Probably more than decade out
Some Profound Questions

- Are there lower bounds on the supply voltages?
- Will reaching these force us to rethink computing models and paradigms?

Shannon-Von Neumann-Landauer Bound for irreversible computing:

Minimum energy/operation = kTln(2)

= $4.10^{-21}$J/bit at room temperature

Many orders of magnitude of opportunity awaiting

MORE LATER ...
In the meantime ...
Truly Energy-Proportional Systems

- Do away with the layers of inefficiency that exist in current systems (any system)

Mostly a software problem, but needs hugely scalable platforms. Means revisiting the business models and the prevailing interfaces.
Truly Energy-Proportional Systems – Incarnations

**SOC Platforms** that effectively exploit every single button to realize proportionality

Even better: Distributed SYSTEMS (computing, networking, storage, IO) that effectively exploit every single button to realize proportionality – Don’t think chip, think system!

Overall impact: proportional to activity – easily factor 10 savings
Always-Optimal Systems

System modules are adaptively biased to adjust to operating, manufacturing and environmental conditions

- Parameters to be measured: temperature, delay, leakage
- Parameters to be controlled: $V_{DD}$, $V_{TH}$ (or $V_{BB}$), clock rate

- Maximum power saving under technology and manufacturing limits
- Inherently improves the robustness of design timing
- Minimum design overhead required over traditional design methodology
Always-Optimal Design

Today’s design: Overdesign & Margin

Tomorrow’s design: Sense & React

Today
- System
  Overspecs & margin
- Subsystem
  Overspecs & margin
- PCB
  Overspecs & margin
- IC’s
  Overspecs & margin
- Circuits

Future
- System
  Sense & React
- Software
  Sense & React
- 3D module
  Sense & React
- IC’s
  Sense & React
- Circuits

A Design Paradigm Shift
Needs built-in controller!

[Courtesy: P. Magarshack, ISSCC 2011]
Enabling Always-Optimal Design

Efficient local supply generation possible at up to ~1W/mm² power density – Allows for integration with < 10% area overhead

Leveraging 2.5D/3D Integration

[Alon, ISSCC11]
Aggressive Deployment (AD) (also known as better-than-worst-case design)

- **Observation:**
  - Worst case conditions rarely encountered in actual operation

- **Remedy:**
  - Operate circuits at lower voltages level than allowed by worst case and deal with the consequences

---

**Example:**
Operate memory at voltages lower than allowed by worst case, and deal with the occasional errors through error-correction

**Distribution ensures that error-rate is low**
Aggressive Deployment

**OVERSCALE voltage and deal with the consequences**

**Example: Razor**

A “pseudo-synchronous” approach to address process variations and power minimization with minimal overhead by combining circuit and architectural techniques

![Graph showing Energy vs Supply Voltage for Optimal Voltage with lines for Total, Processor, and Recovered Energy.]

Adopted by Intel, ARM, AMD, …

[Courtesy: Austin, Blaauw, U Mich]
Sufficient to scale the wall? NOT REALLY!

Maybe adding randomness can help ... 
Going beyond Deterministic Turing Machines

**Probabilistic Turing machine**

From Wikipedia, the free encyclopedia

In computability theory, a **probabilistic Turing machine** is a non-deterministic Turing machine which randomly chooses between the available transitions at each point according to some probability distribution.

In the case of equal probabilities for the transitions, it can be defined as a deterministic Turing machine having an additional "write" instruction where the value of the write is uniformly distributed in the Turing Machine's alphabet (generally, an equal likelihood of writing a '1' or a '0' on to the tape.) Another common reformulation is simply a deterministic Turing machine with an added tape full of random bits called the *random tape*.

As a consequence, a probabilistic Turing machine can (unlike a deterministic Turing Machine) have stochastic results; on a given input and instruction state machine, it may have different run times, or it may not halt at all; further, it may accept an input in one execution and reject the same input in another execution.
The Next Step:
“Non-deterministic digital” or “statistical computing”

Computational engines that, given properties and statistics of input signals and physical implementation, ensure that outputs fall within desired specifications.

- Uses digital encoding
- Non-determinism arises from errors generated in compute modules
- Detection and estimation used to reign in error bounds

\[
y = y_o + \eta + \varepsilon
\]

Distinct error characteristics
The Opportunity: Functional Non-Determinism

App. Domain

- Non-deterministic Computing
- Redundancy/Overdesign

Solution

Required Accuracy

- Efficiency

+ Efficiency

- Required Accuracy
Statistical Computing

Inputs: deterministic or stochastic variables
Outputs: stochastic variables with guaranteed properties (mean, distribution, bounds)

Implementation adds randomness (errors)
System designed such that output metrics are accomplished in spite of randomness of implementation

Requires error models to help design the compensation techniques

Examples: synthesis, classification, modeling, search, recognition
Example: Error-Resilient System Architecture (ERSA)

Typical recognition/mining/synthesis workload (Probabilistic belief propagation, K-means clustering, Bayesian networks)

Cognitive resilience Intolerant to control or higher bit errors

[Courtesy: S. Mitra – Stanford]
### Example: Asymmetric Reliability

- **Highly Reliable (Expensive)**
  - Proper Error Protection
  - Executes Main Thread
  - Assign Worker Threads
  - Reduction
  - Supervise RRCs
  - Timeout check

- **Inexpensive & Unreliable**
  - Without expensive error detection
  - Worker Threads
    - Consists most of the workload
  - Reliable parts
    - Memory Bound Check
    - Restart
RMS on ERSA

Main Thread \((\text{SRC})\)

- Setup
- Work Assignment
- Barrier “Basic” check
- Data Reduction
- Convergence Test

(Work, Memory bounds, Timeout)

Worker thread + bounds check (RRC)

Calculate

Worker thread + bounds check (RRC)

Calculate

Simplistic ERSA inadequate

Convergence filtering heuristics

Convergence damping = Estimation
ERSA Real World App: Image Classifier

- Inferred: a car
- Inferred: not a car

0 Errors / sec / RRC
Correct Result

30K Errors / sec / RRC
90% accurate

- Bayesian Network Inference
  - 90% accuracy enough: cognitive resilience

Incorrect inference: Just one
Algorithmic Noise-Tolerance (ANT)

Combining estimation and detection

- Main Block designed for average case
  - Makes intermittent errors (reduced margins)
- Estimator approximates Main Block output
- Detector compares and replaces
- Assumes algorithmic knowledge for designing efficient estimators

\[ y_a = y_o + \eta \]
\[ y_e = y_o + e \]

[Courtesy: Shanbhag et al, UIUC]
Results: ANT Motion Estimation

2.5X energy-savings

Peak SNR

PSNR variance reduction: 7X

PSNR increase 1.5db
Going one step further ...

- \( P_{\text{avg}}(\text{brain}) = 20 \text{ W} \)
  - (20% of body dissipation, 2% of the weight),
  - Power density: \( \sim 15 \text{ mW/cm}^3 \)
- Nerve cells only 4% of brain volume
- Average neuron density: 70 million/cm\(^3\)
- Computational capacity of the brain: \( 10^{16} \) computations per second*
  - 1-2 fJ per computation (<1 aJ per operation)
  - (Memory capacity: 100k TeraByte)

A VERY energy-efficient computer

Underlying concept: statistical inference

[* R. Kurzweil, “Singularity”]
Neuro-inspired Computing?!

- Robustness in presence of component failure or parametric variations
- Amazing performance with mediocre components
  - Auditory system: can tell difference of time arrival within 10 μs with cells having time constant of 1ms
  - Olfactory system: can discriminate $10^4$-10$^5$ odors with a slight difference of chemical structure with olfactory receptors having broad reception range
- Concepts applicable to many computational problems
- Seamless interaction with analog world
- 2-3 orders more efficient than today’s silicon equivalent
Elements of Neural Processing

<table>
<thead>
<tr>
<th>Sensors</th>
<th>Convergence</th>
<th>Overcompleteness</th>
<th>Associative Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual</td>
<td>Retina</td>
<td>Ganglion Cells/LGN</td>
<td>Primary Visual Cortex (V1)</td>
</tr>
<tr>
<td>Olfactory</td>
<td>Olfactory Epithelium (OE)</td>
<td>Olfactory Bulb (OB)</td>
<td>Primary Olfactory Cortex</td>
</tr>
</tbody>
</table>

Massively parallel, high variability, low SNR computation
Similar to nanoscale devices!
The Lure of Analog Processing

Computational engines that, given the properties and statistics of the input signals and the physical implementation, ensure that the outputs fall within the desired specifications.

**Analog Processing Pro’s and Con’s**
- Single operator performs complex functions
  - Inherently energy efficient
- Improving SNR (resolution) requires more power and/or area
  - Linear or quadratic relationship

\[
P_T = \left( \frac{C_w \cdot SNR \cdot \Delta f}{V_{DD}^2 - \frac{SNR \cdot C_f \cdot \ln(f_h / f_l)}{A_T}} \right)^{1/p}
\]

[Courtesy: R. Sarapeshkar, MIT]
The Lure of Analog Processing

Computational engines that, given the properties and statistics of the input signals and the physical implementation, ensure that the outputs fall within the desired specifications.

Digital Processing Pro’s and Con’s
- Single operator very simple (logic gate)
- Power increases logarithmically with resolution (SNR)
  - 1 extra bit in adder increases SNR with factor 2

\[ P_T = D_p \Delta f \log_2(1 + SNR) \]
The Lure of Analog Processing

Computational engines that, given the properties and statistics of the input signals and the physical implementation, ensure that the outputs fall within the desired specifications.

Digital versus Analog Processing
- Digital shines for high SNR
- Analog is supreme at low SNR

[R. Sarpeshkar, Ultra-Low Power Bioelectronics, ©2010]
Take Aways

- Major reductions in energy/operation not evident in near future
- Major reductions in design margins an interesting proposition
- Computational platforms as dynamically self-adapting and self-regulating feedback systems
- Huge opportunity: most applications do not need huge resolution or deterministic outcomes
- Challenge: needs a rethinking of, applications, algorithms, architectures and platforms, metrics, and ... INSPIRATION
What it means for design methodology?

- The time of deterministic “design-time” optimization is long gone!
  - How to specify, model, analyze and verify systems that dynamically adapt?
- Designs becoming probabilistic engines
  - Input descriptions that capture intended statistical behavior
  - Statistical models of components
  - Algorithm optimization and software generation (a.k.a compilers) so that intended behavior is obtained
- The return of analog!

[Kandinsky, Decisive Pink]