Energy-Efficient Circuit Technologies for Sub-14nm Microprocessors and SoCs: Challenges and Opportunities

IEEE Solid-State Circuits Society Webinar
July 13 2016
Ram K. Krishnamurthy
Senior Principal Engineer
IEEE Fellow & SSCS Distinguished Lecturer
Circuits Research Lab, Intel Labs
Intel Corporation, Hillsboro, OR 97124, USA
ram.krishnamurthy@intel.com

Acknowledgements: Intel Circuits Research Lab, Vivek De, Matt Haycock, Shekhar Borkar, ADR Bangalore Design Lab
Era of Tera-scale Computing

Teraflops of performance operating on Terabytes of data

Performance

Models

3D & Video

Multimedia

Text

KIPS

MIPS

GIPS

TIPS

Dataset Size

Kilobytes

Megabytes

Gigabytes

Terabytes

Financial Analytics

Entertainment, learning and virtual travel

Model-based Apps

Recognition

Mining

Synthesis

Personal Media

Creation and Management

Health

Multi-core

Single-core

Terascale
Internet of Everything (IoE)

Need end-to-end energy efficiency & security
Tera-scale Microprocessors and SoCs

Deliver best user experience under constraints
Moore’s Law scaling

More, better transistors
More cores
Continued benefits from Moore’s Law

14nm Trigate 2014

Source: Intel
Performance/Energy Scaling Trends

Transistor Performance and Power

Transistor Performance vs. Leakage

Innovation Enabled Technology Pipeline

Source: Intel
22nm Interconnects

- M1 to M8 cross-section
- M1-M6 use ultra-low-k ILD and self-aligned vias providing 13-18% capacitance reduction
- Cross-section of integrated MIM capacitor

C. Auth, VLSI Symposium 2012
# Microprocessor Evolution

<table>
<thead>
<tr>
<th></th>
<th>4004 Processor</th>
<th>Westmere-EX Processor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year</strong></td>
<td>1971</td>
<td>2011</td>
</tr>
<tr>
<td><strong>Transistors</strong></td>
<td>2300</td>
<td>2.6 B</td>
</tr>
<tr>
<td><strong>Process</strong></td>
<td>10 µm</td>
<td>32 nm</td>
</tr>
<tr>
<td><strong>Die area</strong></td>
<td>12 mm²</td>
<td>513 mm²</td>
</tr>
</tbody>
</table>

Source: Intel

Die photos not at scale
“Extreme” energy efficiency

10 year goal: ~300X Improvement in energy efficiency
Equal to 20 pJ/FLOPS at the system level
NTV Operation & Energy Efficiency

Frequency reduces almost linearly first, then exponentially
Total power reduces by three to four orders of magnitude

Energy efficiency improves by one order of magnitude at NTV
Energy efficiency reduces in subthreshold operation
Leakage power reduces by two to three orders of magnitude

H. Kaul, R. Krishnamurthy et al, ISSCC 2008
Voltage-frequency range limiters

Vmax/Fmax limiters
- Reliability
- Thermals
- Power delivery

Vmin limiters
- Circuit functional failures
- Soft errors
- Steep frequency roll-off
- Aging

Reliability & functional failures limit range
NTV design techniques

Narrow muxes  No stack height > 2

Robust level converters

Modified Register File Cell (L1$)

Robust Flop Topologies

Multi-corner design optimizations (SCL)

Variation-aware design 2X min Z, 40% lib cells used
NTV operation improves energy efficiency across 45nm-22nm CMOS

H. Kaul, et. al., ISSCC 2009

A. Agarwal, et. al., ISSCC 2010

S. K. Hsu, et. al., ISSCC 2012

ISSCC 2012 Distinguished Paper Award, ESSCIRC 2012 Best Paper Award

NTV operation improves energy efficiency across 45nm-22nm CMOS
**Vector Flip-flops for \( V_{\text{MIN}} \)**

- Min-sized clock inverters shared across adjacent flip-flops \( \Rightarrow \) no clock load increase
- Hold time \( V_{\text{MIN}} \) improves by 175mV

---

**22nm Tri-Gate CMOS Simulation**

\( 0^\circ \text{C-85}^\circ \text{C}, 3\sigma_{\text{systematic}}, 6\sigma_{\text{random}} \)

- \( \Delta \) Flip-flop
- \( \bullet \) Vector Flip-flop

---

\[ S. \text{ Hsu, R. Krishnamurthy et al, ISSCC 2012} \]
ULVS Level Shifter for $V_{\text{MIN}}$

- Decouples CVSL stage from output driver stage and interrupts cross-coupled PMOS devices
- Reduced contention improves $V_{\text{MIN}}$ by 125mV

22nm Tri-Gate CMOS Simulation
$0^\circ\text{C}-85^\circ\text{C}$, $3\sigma_{\text{systematic}}$, $6\sigma_{\text{random}}$

S. Hsu, R. Krishnamurthy et al, ISSCC 2012
DSP functions highly throughput-oriented: Amenable for parallelism/pipelining
⇒ Better power-performance optimization
⇒ Optimal partitioning of tasks between GP processor and dedicated engines
NTV Variable Precision FPU

Accelerator for Smarter Numerics: Variable Precision FP

- Today’s floating-point units wastes energy, time, and storage by using worst-case precision everywhere
- Alternative: use a variable-precision floating point prototype unit with accuracy tracking for multiply-add
- Changing precision (24-bit → 12-bit → 6-bit) on-the-fly can cut energy by 50%

H. Kaul, R. Krishnamurthy et al, ISSCC 2012
Near threshold voltage IA processor

- Technology: 32nm High-K Metal Gate
- Interconnect: 1 Poly, 9 Metal (Cu)
- Transistors: 6 Million (Core)
- Core Area: 2mm²
Power performance measurements

![Diagram showing power performance measurements for 32nm CMOS, 25°C, with various points marked for frequencies and power consumption.]

- 915MHz: 737mW
- 500MHz: 174mW
- 100MHz: 17mW
- 3MHz: 2mW

Frequency (MHz) vs. Total Power (mW)

Logic Vcc / Memory Vcc (V)

- 3MHz: 0.55
- 100MHz: 0.55
- 500MHz: 0.55
- 915MHz: 0.6
Energy efficiency peaks near threshold

![Graph showing energy efficiency and voltage/frequency operating points](image)
NTV and variability

Leakage Comparison

<table>
<thead>
<tr>
<th></th>
<th>Slow</th>
<th>Medium</th>
<th>Fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy/Cycle (pJ)</td>
<td>1.0X</td>
<td>2.5X</td>
<td>7.5X</td>
</tr>
</tbody>
</table>
Interconnect Trends

Number of Metal Layers

Technology Generation (nm)

Al

Cu

500 350 250 180 130 90 65 45 32 22
On-chip Interconnect Trend

- Local interconnects scale with gate delay
- Global interconnects do not keep up with scaling

Source: ITRS
Circuit-Switched On-Chip Interconnects

Improving Efficiency of the On-Die Network

G. Chen, R. Krishnamurthy et al, ISSCC 2014

Hybrid Circuit-/Packet-Switched Network

• Circuit-switching eliminates intra-route data storage
• Packet-switching used only for channel requests
• 2.5X to 3X better efficiency over packet switched network
Environment-aware dynamic adaptation

- Adapt F/V to V/T change → reduce V/T margin
- Adapt F/V to aging → reduce aging margin
Integrated Voltage Regulators: Fine-grain power management

Spatial domain

**Coarse-grain management**
- Same voltage to all cores
- Same frequency for all cores

**Fine-grain management**
- Each core/cluster at optimum voltage
- Each core/cluster at optimum frequency

**TODAY**

- V/F domain interfaces
- Synchronization overhead
- Clock generation/distribution
- Power grid routing
- Optimum V/F for non-cores
- Sub-core clock/leakage gating
- Sub-core V/F domains

**FUTURE**
Nano-AES Hardware Accelerator

- Most popular symmetric-key encryption algorithm
- Conventional 128b AES datapath → large area & power
- Not suitable for ultra-low power wearable systems
22nm CMOS NTV Nano-AES Accelerator

S. Mathew, R. Krishnamurthy et al, 2014 VLSI circuits symposium

<table>
<thead>
<tr>
<th>Process</th>
<th>Encrypt</th>
<th>Decrypt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Die area</td>
<td>0.19mm²</td>
<td></td>
</tr>
<tr>
<td>Gate Count</td>
<td>1947</td>
<td>2090</td>
</tr>
<tr>
<td>Ground-field poly</td>
<td>$x^4 + x^3 + 1$</td>
<td>$x^4 + x + 1$</td>
</tr>
<tr>
<td>Extension-field poly</td>
<td>$x^2 + 6x + 9$</td>
<td>$x^2 + 2x + E$</td>
</tr>
<tr>
<td>Area (µm²)</td>
<td>2200</td>
<td>2736</td>
</tr>
</tbody>
</table>

Industry-leading energy efficiency of 289Gbps/Watt at 340mV!
• \[11x\] higher energy-efficiency than previously-reported measurements
AES symmetric-key crypto accelerator

S. Mathew, R. Krishnamurthy et al, 2010 VLSI circuits symposium

[Diagram of AES encryption/decryption process]
All-digital random number generator

S. Mathew, R. Krishnamurthy et al, 2015 ESSCIRC
S. Mathew, R. Krishnamurthy et al, 2010 VLSI circuits symposium

Scalable & PVT variation tolerant
Physically unclonable function (PUF)

S. Mathew, R. Krishnamurthy et al, 2014 ISSCC
Neuromorphic computing
“Extreme” efficiency research

System-Wide Breakthroughs Needed Across the Board

- Extreme Energy Efficiency
- Fine-Grain Power Management
- Efficient Memory Subsystem
- Self-Aware Computing Operation
- Programming for Extreme Parallelism
Legal Disclaimer

This presentation contains the general insights and opinions of Intel Corporation (Intel).

- This presentation is provided for **informational purposes only** and is not to be relied upon for any other purpose. Intel makes no representations or warranties regarding the accuracy or completeness of the information in this presentation. Intel accepts no duty to update this presentation based on more current information. Intel is not liable for any damages, direct or indirect, consequential or otherwise, that may arise, directly or indirectly, from the use or misuse of the information in this presentation. Intel retains all rights to the Presentation, including any patent, trademark, trade secret, copyright, trade dress, mask works, or any other intellectual property rights. The provision of the Presentation does not constitute the grant or license of any such rights by Intel. Provision of the Presentation shall not be construed to constitute advice or consultation. Intel does not provide, and is not providing, any technical, legal, regulatory or compliance advice. Nor does Intel make any representation or warranty with respect to the effectiveness of any information contained herein.

- Intel may make changes to specifications and product descriptions at any time, without notice.
- Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations. Intel does not control or audit the design or implementation of third party benchmarks or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmarks are reported and confirm whether the referenced benchmarks are accurate and reflect performance of systems available for purchase.
- Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. See www.intel.com/products/processor_number for details. Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Intel Virtualization Technology requires a computer system with a processor, chipset, BIOS, virtual machine monitor (VMM) and applications enabled for virtualization technology. Functionality, performance or other virtualization technology benefits will vary depending on hardware and software configurations. Virtualization technology-enabled BIOS and VMM applications are currently in development.
- Intel® Turbo Boost Technology requires a Platform with a processor with Intel Turbo Boost Technology capability. Intel Turbo Boost Technology performance varies depending on hardware, software and overall system configuration. For more information, see http://www.intel.com/technology/turboboost.
- No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology (Intel® TXT) is a security technology under development by Intel and requires for operation a computer system with Intel® Virtualization Technology, a Intel® Trusted Execution Technology-enabled Intel processor, chipset, BIOS, Authenticated Code Modules, and an Intel or other Intel® Trusted Execution Technology compatible measured virtual machine monitor. In addition, Intel® Trusted Execution Technology requires the system to contain a TPMv1.2 as defined by the Trusted Computing Group and specific software for some uses. 64-bit computing on Intel architecture requires a computer system with a processor, chipset, BIOS, operating system, device drivers and applications enabled for Intel® 64 architecture. Performance will vary depending on your hardware and software configurations. Consult with your system vendor for more information.
- Intel, Intel Xeon, Intel Core microarchitecture, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. © 2008 Standard Performance Evaluation Corporation (SPEC) logo is reprinted with permission.
- Other names and brands may be claimed as the property of others.