2023 Texas Symposium on Computing with Emerging Technologies (ComET)  
University of Texas at Dallas

Limits of CMOS and Prospects for Adiabatic/Reversible CMOS

Monday, October 30th, 2023

Michael P. Frank, Center for Computing Research, SNL  
& Alexander J. Edwards, University of Texas at Dallas

with collaborators: Brian D. Tierney (SNL), Joseph Friedman (UT Dallas)

Approved for public release
Abstract (hide during talk)

Limits of CMOS and Prospects for Adiabatic/Reversible CMOS

The energy efficiency of conventional CMOS logic is fast approaching practical limits which ultimately arise from fundamental physical considerations. The minimum typical logic signal energy is projected to bottom out at around 0.2 fJ (1.25 keV) by around 2030 on the IRDS roadmap. This will exacerbate the tension between achievable device densities (which will continue to increase as the industry moves towards 3D VLSI techniques in which multiple “tiers” of active devices can be integrated within a single fabrication process), versus the need for the power dissipation density within chip packages to remain manageable. Effectively, these constraints will result in the potentially-available device-count resources becoming increasingly massively underutilized in practical chip designs, compounding the issues of “dark silicon” that already exist today.

The principles of fully adiabatic switching offer an alternative, relatively little-explored technology development path for CMOS which can mitigate these problems, allowing the energy dissipation per switching event to continue being reduced as technology advances, thereby improving achievable throughput within package-level power dissipation constraints, and permitting maximal utilization of the affordable device counts within a given package. The potential advantages of this approach continue to increase as manufacturing processes continue to advance and additional tiers of active logic are fabricated within a die, and/or multiple die or chiplets are stacked up in 3D within a package, with the ultimate limits of digital performance per unit power consumption or package area still being far away, but only if these methods are leveraged.

In this talk, we will review how the practical limits on the efficiency of conventional CMOS arise from fundamental physical considerations, and discuss how adiabatic switching principles, when applied properly, can allow us to circumvent these limits. Then we will give a preview of preliminary results from our work in progress on analyzing the maximum boosts in raw throughput density, as a function of per-die power dissipation density, that can theoretically be achieved through utilizing the principles of fully adiabatic switching. Early results suggest low-level efficiency and throughput density can be boosted by up to nearly 400× using these methods vs. conventional CMOS, assuming standard specifications for off-state leakage conductance per unit device width.
Contributors to our Reversible Computing research program

Full group of recent staff engaged at Sandia:

- **Michael Frank** (Cognitive & Emerging Computing)
- **Reza Arghavani** (Regional Security & Analysis)
- **Robert Brocato** (RF Microsystems) – now retired
- **David Henry** (MESA Hetero-Integration)
- **Rupert Lewis** (Quantum Phenomena)
  - Terence “Terry” Michael Bretz-Sullivan
- **Nancy Missert** (Nanoscale Sciences) – now retired
  - **Matt Wolak** (now at Northrop-Grumman)
- **Brian Tierney** (Rad Hard CMOS Technology)

Thanks are also due to the following colleagues & external research collaborators:

- **Rudro Biswas** (Purdue)
  - Students Dewan Woods & Rishabh Khare
- **Tom Conte** (Georgia Tech/CRNCH)
  - Anirudh Jain, Gibran Essa, Austin Adams
- **Erik DeBenedictis** (Sandia ➔ Zettaflops, LLC)
- **Hannah Earley** (Cambridge U. ➔ startup)
- **Joseph Friedman** (UT Dallas)
- **David Guéry-Odelin** (Toulouse U.)
- **Steve Kaplan** (independent contractor)
- **Kevin Osborn** (LPS/JQI)
  - With Liuqi Yu, Ryan Clarke, Han Cai
- **Karpur Shukla** (CMU ➔ Flame U. ➔ Brown U.)
  - Prev. in Prof. Jimmy Xu’s Lab for Emerging Techs.
- **FAMU-FSU College of Engineering**:
  - Sastry Pamidi (ECE Chair) & Jerris Hooker (Instructor)
  - 2019-20 students:
    - Frank Allen, Oscar L. Corces, James Hardy, Fadi Matloob
  - 2020-21 students:
    - Marshal Nachreiner, Samuel Perlman, Donovan Sharp, Jesus Sosa

Thanks are due to Sandia’s LDRD program, DOE’s ASC and SCGSR programs, and the DoD/ARO ACI (Advanced Computing Initiative) for their support of our research!
1. We are now only \( \sim 10 \times \) away from ultimate limits on the (low-level) energy efficiency of conventional CMOS!
   - Irrevocable fundamental device-level energy limits imply much closer limits for practical logic!
   - The practical limits on 8-bit arithmetic \( \approx \) the energy used by the brain per synapse firing (!)
   - Leads to severe limits on scaling of performance density (ops/sec/area) given cooling constraints.

2. **Fully adiabatic switching** provides a path to circumvent this limit in digital CMOS!
   - Principles of adiabatic switching applied to CMOS suggest \( >100 \times \) raw efficiency boosts are possible
   - Most of the dynamic power in the circuit can be resonantly recirculated, and not dissipated to heat
   - Permits effective utilization of more active gates per die, more layers of active processing per package

3. Focus of present work: Analyze raw throughput density boost from fully adiabatic switching for future CMOS as a function of (per-die) power dissipation density.
   - Utilize approximate device models based on IRDS roadmap data for six process nodes (2022–2037).
   - Consider both conventional and adiabatic switching, at both nominal and optimized voltage levels.
   - Optimize average density of active gates (per die), logic swing voltage, and switching frequency for maximum throughput density

4. Conclusions
   - Substantial (orders of magnitude) further gains in the raw efficiency of general digital tech beyond the limits of conventional digital logic are potentially available in CMOS…
     - but only if the principles of adiabatic switching and reversible computing are aggressively applied!
I. Limits on the Energy Efficiency of Conventional CMOS Technology
## A Tale of Two Systems
(Note both are DOE supercomputers that each led the TOP500 list in their day)

<table>
<thead>
<tr>
<th>Then:</th>
<th>Now:</th>
<th>Comparison:</th>
<th>Ann. Chg.:</th>
<th>Per Decade:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year:</td>
<td>1997</td>
<td>2022</td>
<td>+ 25 years</td>
<td>+1 year</td>
</tr>
<tr>
<td>System Name:</td>
<td>ASCI Red</td>
<td>Frontier</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location:</td>
<td>Sandia (NM)</td>
<td>Oak Ridge (TN)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perf. (max. sust.):</td>
<td>1.068 Tflop/s</td>
<td>1.102 Eflop/s</td>
<td>Perf. 1.032 million ×</td>
<td>+ 74.0%</td>
</tr>
<tr>
<td>Power draw:</td>
<td>850 kW</td>
<td>21.1 MW</td>
<td>Power ~25 ×</td>
<td>+ 13.7%</td>
</tr>
<tr>
<td>Efficiency:</td>
<td>1.256 Mflops/W</td>
<td>52.23 Gflops/W</td>
<td>Efficiency 41,570 ×</td>
<td>+ 53.0%</td>
</tr>
<tr>
<td>Process Tech.:</td>
<td>250 nm</td>
<td>“3 nm”</td>
<td>Density ~6,900×</td>
<td>+ 42.5%</td>
</tr>
<tr>
<td>Min. Gate Energy:</td>
<td>~ 1 fJ</td>
<td>~5 aJ</td>
<td>Device Effic. 200 ×</td>
<td>+ 23.6%</td>
</tr>
<tr>
<td>Arch. Eff. (arb. units):</td>
<td>1</td>
<td>207.8</td>
<td>Arch. Effic. ~208 ×</td>
<td>+ 23.8%</td>
</tr>
</tbody>
</table>

- Note that over the last quarter-century, effic. of low-level device tech. & system architectures improved roughly in sync
  - Both improved by ~200×/25yr. = ~8.3×/10yr. on average over this period
Rates of Performance Improvement Have Not Been Uniform!

There was a clear change in the slope of the system-level performance growth trendline at the start of 2013!

- Prior to 2013, average system performance among TOP500 supercomputers improved at a fairly steady rate of $\sim 460\times$/decade.
- Starting in 2013, performance growth declined to a much slower rate of $\sim 28$/decade.

This may be attributed to a delayed system-level response to the plateauing of clock speeds that occurred in $\sim 2005$

- After a few years, chip architects & system integrators ran out of other tricks to maintain system performance growth rate
- The ITRS roadmap framers deliberately slowed the pace for forward-looking system performance targets in response
Looking forward now...

International Roadmap for Devices & Systems (IRDS)

Plus additional chapters and white papers...
# The “More Moore” chapter – specifies technology node targets

<table>
<thead>
<tr>
<th>YEAR OF PRODUCTION</th>
<th>2022</th>
<th>2025</th>
<th>2028</th>
<th>2031</th>
<th>2034</th>
<th>2037</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logic industry &quot;Node Range&quot; Labeling</td>
<td>3nm*</td>
<td>2nm*</td>
<td>1.5nm*</td>
<td>1.0nm eq*</td>
<td>0.7nm eq*</td>
<td>0.5nm eq*</td>
</tr>
<tr>
<td>Fine-pitch 3D integration scheme</td>
<td>Stacking</td>
<td>Stacking</td>
<td>Stacking</td>
<td>3DVLSi</td>
<td>3DVLSi</td>
<td>3DVLSi</td>
</tr>
<tr>
<td>Logic device structure options</td>
<td>finFET</td>
<td>LGAA</td>
<td>LGAA</td>
<td>LGAA-3D</td>
<td>LGAA-3D</td>
<td>LGAA-3D</td>
</tr>
<tr>
<td>Platform device for logic</td>
<td>finFET</td>
<td>LGAA</td>
<td>LGAA</td>
<td>LGAA-3D</td>
<td>LGAA-3D</td>
<td>LGAA-3D</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2022</th>
<th>2025</th>
<th>2028</th>
<th>2031</th>
<th>2034</th>
<th>2037</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vdd (V)</td>
<td>0.70</td>
<td>0.65</td>
<td>0.65</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>Gate length (nm)</td>
<td>16</td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Number of stacked tiers [1]</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Number of stacked nanosheets in logic device [1]</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Number of stacked nanosheets in SRAM device [1]</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Maximum number of stacked nanosheets in one device [1]</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>16</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Digital block area scaling</td>
<td>1.00</td>
<td>0.74</td>
<td>0.55</td>
<td>0.26</td>
<td>0.13</td>
<td>0.08</td>
</tr>
<tr>
<td>Digital block energy scaling</td>
<td>1.00</td>
<td>0.81</td>
<td>0.72</td>
<td>0.56</td>
<td>0.50</td>
<td>0.49</td>
</tr>
<tr>
<td>#MAC units in SoC - based on integration capacity</td>
<td>8192</td>
<td>11038</td>
<td>14980</td>
<td>30966</td>
<td>65191</td>
<td>108652</td>
</tr>
<tr>
<td>Cell height (nm) - HD</td>
<td>144</td>
<td>114</td>
<td>90</td>
<td>80</td>
<td>80</td>
<td>72</td>
</tr>
<tr>
<td>CPU frequency (GHz)</td>
<td>3.18</td>
<td>3.28</td>
<td>3.36</td>
<td>3.42</td>
<td>3.47</td>
<td>3.50</td>
</tr>
<tr>
<td>CPU frequency at constant power density (GHz)</td>
<td>3.18</td>
<td>3.17</td>
<td>2.79</td>
<td>1.49</td>
<td>0.71</td>
<td>0.44</td>
</tr>
<tr>
<td>Power density scaling</td>
<td>1.00</td>
<td>1.03</td>
<td>1.20</td>
<td>2.29</td>
<td>4.85</td>
<td>7.99</td>
</tr>
<tr>
<td>TOPS/mm² scaling</td>
<td>1.00</td>
<td>1.39</td>
<td>1.93</td>
<td>4.07</td>
<td>8.68</td>
<td>14.62</td>
</tr>
<tr>
<td>TOPS/W scaling</td>
<td>1.00</td>
<td>1.23</td>
<td>1.39</td>
<td>1.79</td>
<td>1.99</td>
<td>2.03</td>
</tr>
<tr>
<td>TOPS/mm² * TOPS/W</td>
<td>1.00</td>
<td>1.71</td>
<td>2.70</td>
<td>7.29</td>
<td>17.24</td>
<td>29.72</td>
</tr>
</tbody>
</table>
The Modern Transistor: Nanosheet Gate-All-Around (GAA) FET

Example process: IBM’s “2 nm” process, announced in 2021, IRDS target date 2025

Nanosheet width: 15-70 nm (here 40 nm)
Nanosheet thickness: ~5-7 nm
Gate length: 12 nm
The end of (energy efficiency improvements in) conventional CMOS is nigh!

Only ~2× remaining on the roadmap!
Only ~8× to the thermal noise limit!

IRDS 2022 "More Moore" Roadmap Energy Targets

- Logic node, nominal voltages
- Roadmap $\frac{1}{2}CV^2$, nominal voltages

Only ~2× improvement left per the roadmap!

Maybe another 4× or so left till thermal noise is an issue!

Thermal noise danger zone!!
An Interesting Comparison… Who Would Win?

Nvidia H100 SXM GPU

| FP8 Perf.: | 3.96 Pflop/s |
| Max power: | 700 W |
| Energy/FP8: | 253 fJ |

= 52.6 million kT
(assuming 75°C operating temp.)

If a synapse firing is roughly comparable computationally to an FP8 operation (e.g., add synapse’s weight into neuron’s activation), then an H-100 is only ~8× less energy efficient than the human brain!

(Note: the below are very rough estimates only!)

Est. #neurons/brain: ~100 billion
Est. synapses/neuron: ~10,000
Est. synapses/brain: ~1 quadrillion
Average neuron fires: ~0.7x/sec.
Aggregate synapse firings: ~700 trillion/sec.
Brain power consumption: ~20 W
Energy per synapse firing: ~28.6 fJ
≈ 6.67 million kT
(assuming 98.6°F operating temp.)

The limits of CMOS vs. human brain efficiency are about the same!
Coincidence? 🤔 Or not?
Thermal fluctuations and the Boltzmann distribution

- **Discovery of thermal fluctuations**
  - Ludwig Boltzmann (1868): Formulated the statistical foundations for understanding thermal phenomena, including the Boltzmann distribution.
  - Albert Einstein (1905): Theoretical explanation linking Brownian motion to thermal fluctuations.

- Boltzmann’s derivation of probability distribution over subsystem energies above a ground state
  - Showed all systems in thermal equilibrium experience random energy fluctuations obeying what’s now called the Boltzmann distribution:

\[ f(E) = Ae^{-E/kT} \]

\[ P(E) \propto e^{-E/kT} \]

Thermal fluctuations in CMOS

- Thermal fluctuations are the fundamental phenomenon that sets the practical limits of CMOS energy efficiency!
- Subthreshold currents are controlled by thermionic emission – thermal excitation of electrons onto potential energy barriers.

In subthreshold:

\[ I_{ds} \propto g_{ds} \propto \exp \left( \frac{V_{ch}}{kT/q} \right) \]

\[ V_{on/off} \leq \exp \left( \frac{V_{dd}}{kT/q} \right) \]

\[ V_{dd} \geq \frac{kT}{q} \ln \left( \frac{g_{on/off}}{g_{ds(V_{gs} = 0)}} \right) \]

\[ E_{sw} \geq kT \ln \left( \frac{g_{on/off}}{g_{ds(V_{gs} = 0)}} \right) \]

(per electron in channel!)

- ~30 kT/channel in 2028
Energy Efficiency limits Throughput Density…

The aggregate computational throughput (ops/sec) per unit area of CMOS is already primarily limited by power dissipation constraints today – and on the conventional path, this problem will grow far worse in the future…

- Note that on the roadmap, efficiency is only improving $2 \times$ by 2037, and they project that for throughput density to increase by the maximum of $\sim 14.6 \times$, power dissipation density would have to increase by $\sim 8 \times$!
- Imagine trying to cool a GPU chip of fixed area that now dissipates 5,600 W instead of 700 W!
- Or, if what we want is to keep power density constant, processor clock speeds would have to fall $\sim 7.2 \times$—e.g. a 3.18 GHz core must be slowed to 440 MHz.
- And then, throughput density only increases in proportion to efficiency ($i.e.$, by only $2.03 \times$).

Through improved efficiency, adiabatic switching can give us a more favorable scaling of throughput density as we go down the roadmap!
- And together with die stacking, can increase throughput density even further!

![Graph showing TOPS/mm² and TOPS/W scaling](image)

(Source: IRDS '22 More Moore chapter)
Cooling system designs are already starting to get insane as it is…

*E.g.*, Cerebras WSE-2 is the largest, highest-performing single AI chip today… BUT it uses up to 23 kW!

- And just look at what-all that requires in terms of cooling hardware *already*… *(How would you boost this another 8x → 184 kW?)*
II. Adiabatic Switching as a Path forward for CMOS Efficiency
Moving Beyond the Thermal Noise Limit…

Thermal noise sets a strict lower bound on gate switching energy, but…

- There’s no fundamental reason why this energy has to be **dissipated** to heat!

**Adiabatic switching** provides a means to **recover** most of the gate energy.
- Pioneered by MIT, CalTech, Xerox PARC, USC/ISI, Rutgers in late '70s-early 90s.

Based on **gradual** logic transitions controlled by an **AC** waveform
- As opposed to **abrupt** switching between **DC** supplies in conventional logic.

Ordinary CMOS dissipates $\frac{1}{2}CV^2$ in each (sudden) switching event…
- Consequence of $Q = CV$ charge delivered from voltage $V \rightarrow$ later returned to 0V.

In **adiabatic** CMOS, we instead deliver charge in a gradual, steady flow…
- We can think of the source as being constant **current** instead of constant **voltage**.
- Can approximate constant current source with a ~linear **voltage ramp** over time $t$.
- Because the charge transfer is more gradual, the voltage drop over the charging path is smaller, and so the energy dissipated during the charge transfer is smaller.

**We basically can make the energy dissipation as small as we want.**
- Down to a lower limit set by leakage.
To approach ideal reversible computing in CMOS…

We must aggressively eliminate all sources of non-adiabatic dissipation, including:

- Diodes in charging path, “sparking,” “squelching,”
  - Eliminated by “truly, fully adiabatic” design. (E.g., CRL, 2LAL).
  - Can suffice to get down to a few aJ (10s of eV) even before voltage optimization!

- Voltage level mismatches that dynamically arise on floating nodes before reconnection.
  - Eliminated by static, “perfectly adiabatic” design. (E.g., S2LAL).

We must also aggressively minimize standby power dissipation from leakage, including:

- Subthreshold channel currents.
  - Ultra-low-\(T\) (e.g., 4K) operation helps with this.

- Tunneling through gate oxide.
  - E.g., use thicker gate oxides.

Note: (Conditional) logical reversibility follows from perfect adiabaticity.

---

See Frank et al. “Exploring the Ultimate Limits of Adiabatic CMOS”, 38th IEEE Int'l Conf. on Computer Design (ICCD'20), 10.1109/ICCD50377.2020.00018

“Perfectly Adiabatic” Reversible Computing in CMOS

---

Shift Register Structure and Timing in 2LAL

---

Shift Register Structure and Timing in S2LAL

---

(512x512)
Simulation Results from the “Adiabatic Circuits Feasibility Study” Efforts at Sandia, funded via NSCI (2017-present)

Created schematic-level fully-adiabatic designs for Sandia’s in-house (MESA) processes, including:
- Older, 350 nm process (blue curve)
  - FET widths = 800 nm
- Newer, 180 nm process (orange, green curves)
  - FET widths = 480 nm

Plotted energy dissipation per-transistor in shift registers at 50% activity factor (alternating 0/1)
- 2LAL (blue, orange curves)
- S2LAL (green curve)

In all of these Cadence/Spectre simulations,
- We assumed a 10 fF parasitic wiring capacitance on each interconnect node.
- Logic supply ($V_{dd}$) voltages were taken at the processes’ nominal values.
  - 3.3V for the 350nm process; 1.8V in the 180nm process.

We expect these results could be significantly improved by exploring the parameter space over possible values of $V_{dd}$.

Note this is ~14,300× smaller vs. brain dissipation per synapse firing!
Trapezoidal Resonators via Fourier Decomposition

We can efficiently generate non-sinusoidal waves using harmonic resonators!

° Consider the ideal trapezoidal waveform shown below

Note, relative to mid-level crossing, trapezoidal waveform is an odd function

° \( \therefore \) Spectrum includes only odd-numbered harmonics \( f, 3f, 5f, \ldots \)

Six-component Fourier series expansion for 2LAL waveform is shown below

° Maximum error with \( 11f \) frequency cutoff is < 1.7% of \( V_{dd} \)

\[
v_{f6}(t) = V_{DD} \left[ \frac{1}{2} + \frac{4\sqrt{2}}{\pi^2} \left( \frac{\sin(\omega t)}{3} \right) + \frac{\sin(3\omega t)}{9} - \frac{\sin(5\omega t)}{25} \right]
\]

Trapezoidal Resonator Circuit Design Concept Invented at Sandia

Work done in our project, 2017–2021
- Patent was issued in 2023

Approach uses a transformer-coupled series of $LC$ tank circuits
- Subcircuit resonant frequencies can be tuned by trimming capacitor sizes
- Relative phases and amplitudes of harmonics are set using transformer winding directions & turn ratios

Resonator $Q$ value was $\sim 3,000$ in simulations with a simple model load
- More fine-grained simulations with a more detailed load model needed
- Prototype development including 3D integration and packaging needed
III. Raw Throughput Density Boosts Achievable via Adiabatic Switching
Analysis of Throughput Density Boost from Adiabatic Switching

Overall approach:

1. For each roadmap year,
   - Estimate a rough device model (giving on-conductance vs. operating voltage) based on roadmap data, and then do:
2. For various power density constraints,
   - (where we explored the 4-OOM range from 10 mW/cm\(^2\) to 100 W/cm\(^2\)), do the following:
3. For various possible logic swing (\(V_{dd}\)) voltages up to the nominal roadmap level,
   - Consider a unit consisting of a generic logic gate and load, as per the roadmap, and do the following:
4. If off-stage leakage power at maximum gate density exceeds the power density constraint,
   - Decrease gate density below maximum until leakage is no greater than 10% of constraint (for conventional logic) or 50% of constraint (for adiabatic)
   - Note that keeping a relatively lower gate density in the leakage-constrained regime does not penalize conventional logic (relative to adiabatic), since its throughput is limited by switching power, not by maximum switching speed anyway
   - Note: once we are in the leakage-dominated regime, adiabatic scales no better with power density than conventional
5. Select the switching frequency such that the power dissipation from active switching plus the leakage power meets (but does not exceed) the power density constraint.
   - Note that the formula for the optimum frequency differs for the adiabatic vs. conventional cases \(\rightarrow\) different scaling!
   - This ends up allowing the adiabatic case to switch at a higher frequency than conventional logic within the constraint!
6. Calculate and plot the raw switching throughput density (logic node switching events per unit time per unit area) from the gate density and switching frequency.
   - Compare these four cases:
     (a) standard-voltage conventional, (b) optimized-voltage conventional,
     (c) standard-voltage adiabatic, and (d) optimized-voltage adiabatic.

The next four slides show preliminary results from our analysis. (Pending refinement.)
Standard-Voltage Conventional Switching

Colors show roadmap years (red = 2022 through magenta = 2037)

Here, we maintain leakage power at no more than 10% of total power by decreasing average gate density as needed.

With conventional switching at standard voltages, throughput falls ∝ power density, as expected – since energy dissipation per switching event is a constant.

- Note that at max density of active gates, switching frequency can be no more than ~1 MHz/W in '22!

Note also that throughput improves by only about 2× between 2022 and 2037!
- Because, see slide 14.
Voltage-Optimized Conventional Switching

Note optimal voltages for maximum throughput density start near threshold, and trend subthreshold at lower power levels.
- End up at roughly $\frac{1}{2}$ of threshold level.

Because of low $V_{dd}$, leakage power is greatly reduced, and doesn’t start to limit max gate density until very low power density levels.

Maximum frequency at max gate density also improves vs. higher-$V_{dd}$, and more so as the power limit & switching voltage decreases.
- $\sim 24.6 \times$ throughput boost at low power per die
Here, we make sure that leakage is no more than 50% of total power
- Because that is the point of theoretical maximum efficiency for adiabatic switching

Note adiabatic frequency vs. power curves are ½ as steep as standard-voltage conventional.
- Increasing adiabatic advantage at low per-die power densities!

Adiabatic gates at standard voltages are more energy efficient / can switch more frequently (at high device densities) than conventional gates at optimized voltages!
- Throughput boosts as high as 21.3×! (@0.18W/cm² in 2022)
Voltage-Optimized Adiabatic Switching

This time, the optimal voltages end up near-threshold but not significantly subthreshold

- Note this improves noise immunity vs. optimized conventional CMOS

Adiabatic scaling advantage extends farther before limited by leakage.
- Maximum boost vs. conventional CMOS is now \(104\times\) (in 2028) at low power-per-die levels
Here, we are running each technology variation at the voltage & frequency that gives it ~max. throughput density/W (@ 0.01 W/die).

Suggests that even beyond the end of the roadmap, we can continue improving energy efficiency by up to another ~2,400× assuming noise isn’t yet limiting (at channel energy ~27 kT).

At the same voltage, conventional CMOS would be only ~6x lower than end-of-roadmap with standard voltages!

Adiabatic beats conventional by ~405× at opt. adia. voltage (0.245 V) if it’s achievable.
What’s wrong with standard voltage scaling?

Note that, for maximum throughput with conv. switching, we would have to push channel energies far down into the “thermal noise danger zone!”

And note that it still is not as efficient as adiabatic switching, even then!
IV. Conclusion
Conclusion & Next Steps

Preliminary conclusions from the present study to date:

◦ Conventional CMOS is fast approaching fundamental limits from thermal noise!
  ◦ Only $\sim2\text{–}12\times$ estimated efficiency improvement remaining till end of roadmap in early-mid 2030s!
  ◦ Depending on how far operating voltages can effectively be lowered below nominal $V_{dd}$ levels.
  ◦ Questions arise about how much farther beyond this we could realistically proceed with conventional switching even if trying to utilize aggressive subthreshold logic levels.
  ◦ Fluctuations in channel energy could significantly impact device function on short timescales

◦ But, adiabatic switching offers a potential workaround for this problem!
  ◦ Raw throughput density (logic switching events/time/area) benefits by up to $\sim100\times$ vs. end of conventional CMOS (even including subthreshold CMOS!), or $\sim400\times$ if comparing @ threshold.
  ◦ And this is before even attempting to optimize device sizing or fab process
  ◦ Not yet accounting for architectural overheads of adiabatic/reversible design, though…

Some appropriate next steps would include:

◦ Make our current crude device models somewhat more realistic, refine analysis
  ◦ Should really include gate leakage! (Presently not included in our simple device model.)
  ◦ Possibly upgrade analysis to include effect of optimizing device widths for adiabatic case
  ◦ Analyze tradeoffs and additional gains available through further minimizing device leakage.

◦ Do some much more detailed circuit-level simulations
  ◦ E.g., integrate resonant oscillator designs driving the logic

◦ Begin a more detailed accounting of well-optimized architectural overheads for example applications
  ◦ E.g., a matrix multiplier core for AI applications