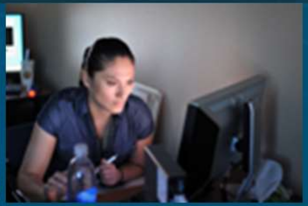
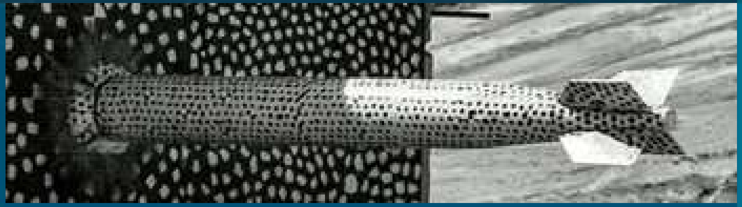
SAND2023-11781C

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

Sandia National Laboratories

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

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

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)

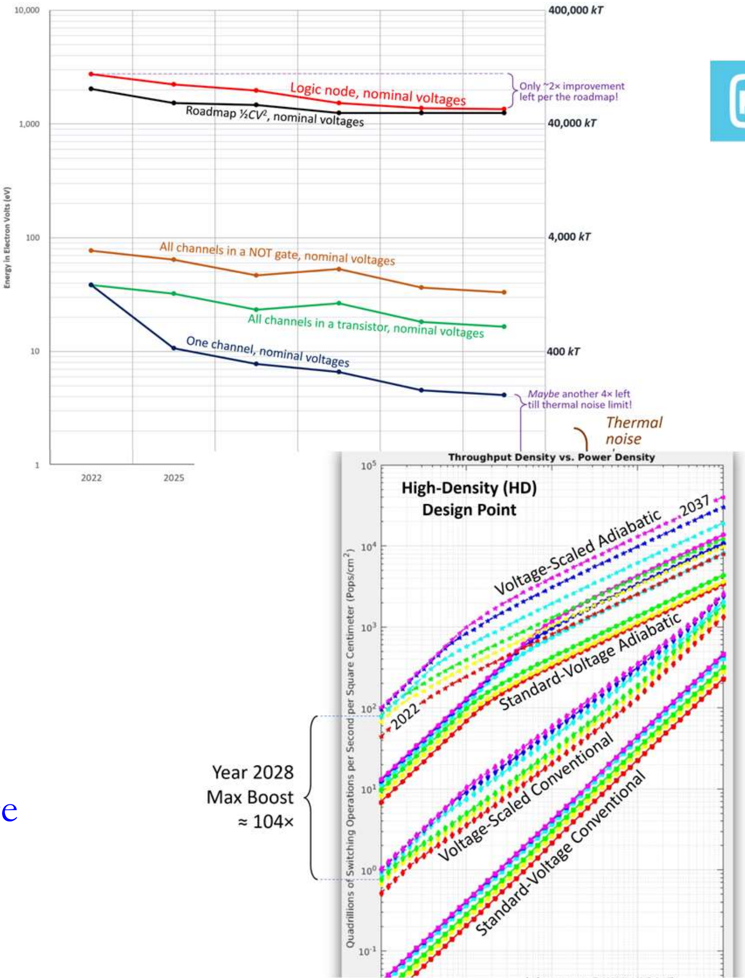
Thanks are due to Sandia's LDRD program, DOE's ASC and SCGSR programs, and the DoD/ARO ACI (Advanced Computing Initiative) for their support of our research!

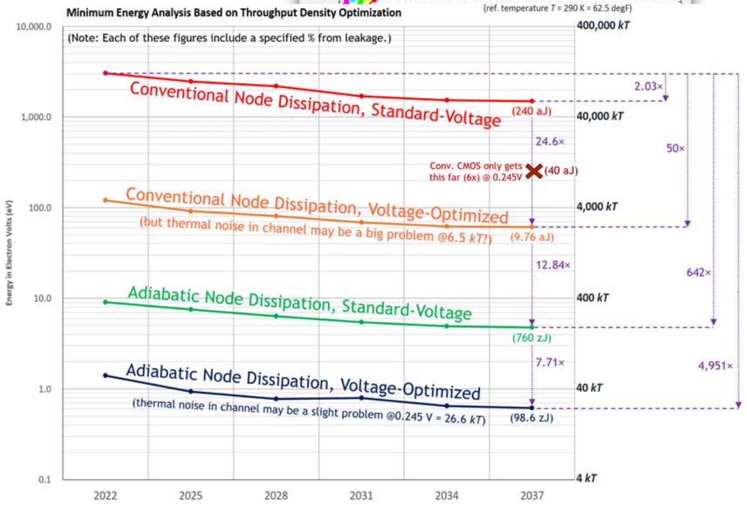
- Thanks are also due to the following colleagues & external research collaborators:
  - Rudro Biswas (Purdue)
    - Students Dewan Woods & Rishabh Khare
  - Tom Conte (Georgia Tech/CRNCH)
    - Anirudh Jain, Gibran Essa, Austin Adams
  - Erik DeBenedictis (Sandia → Zettaflops, LLC)
  - Hannah Earley (Cambridge U. → startup)
  - Joseph Friedman (UT Dallas)
    - with A. Edwards, F. Garcia-Sanchez, X. Hu, J.A.C. Incorvia, A. Paler, B.W. Walker, P. Zhou
  - 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

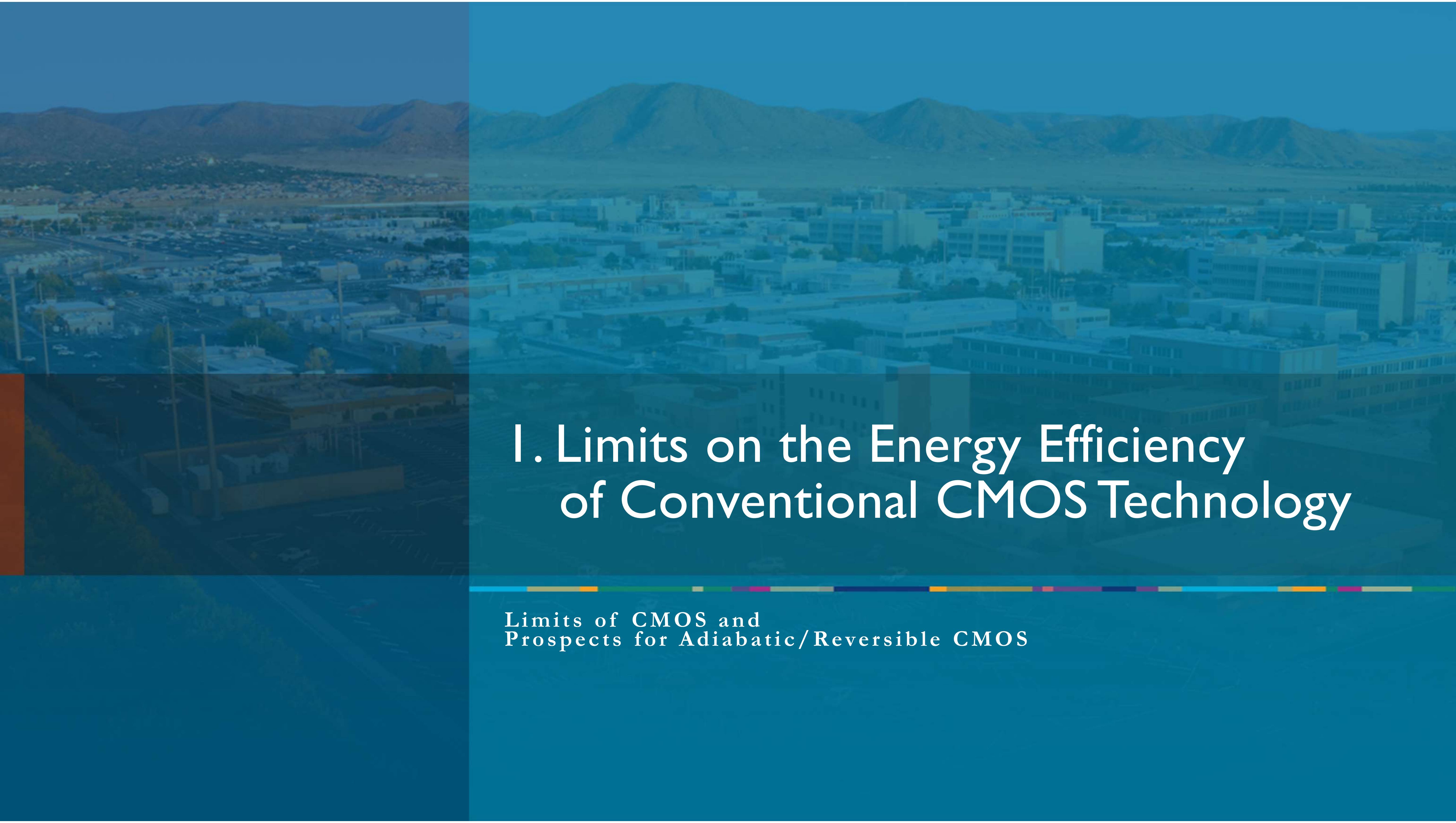
Talk Abstract/Outline

Limits of CMOS and Prospects for Adiabatic/Reversible CMOS

1. We are now only **~10× 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 ≈ 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× 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!







1. Limits on the Energy Efficiency of Conventional CMOS Technology

Limits of CMOS and
Prospects for Adiabatic/Reversible CMOS

A Tale of Two Systems

(Note both are DOE supercomputers that each led the TOP500 list in their day)

	Then:	Now:	Comparison:	Ann. Chg.:	Per Decade:
Year:	1997	2022	+ 25 years	+1 year	+10 years
System Name:	ASCI Red	Frontier			
Location:	Sandia (NM)	Oak Ridge (TN)			
Perf. (max. sust.):	1.068 Tflop/s	1.102 Eflop/s	Perf. 1.032 million ×	+ 74.0%	Perf. 254 ×
Power draw:	850 kW	21.1 MW	Power ~25 ×	+ 13.7%	Power 3.6 ×
Efficiency:	1.256 Mflops/W	52.23 Gflops/W	Efficiency 41,570 ×	+ 53.0%	Eff. 70.4 ×
Process Tech.:	250 nm	“3 nm”	Density ~6,900×	+ 42.5%	Dens. 34.4 ×
Min. Gate Energy:	~ 1 fJ	~5 aJ	Device Effic. 200 ×	+ 23.6%	Dev. Eff. 8.3 ×
Arch. Eff. (arb. units):	1	207.8	Arch. Effic. ~208 ×	+ 23.8%	Arch. Eff. 8.4 ×

- 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!

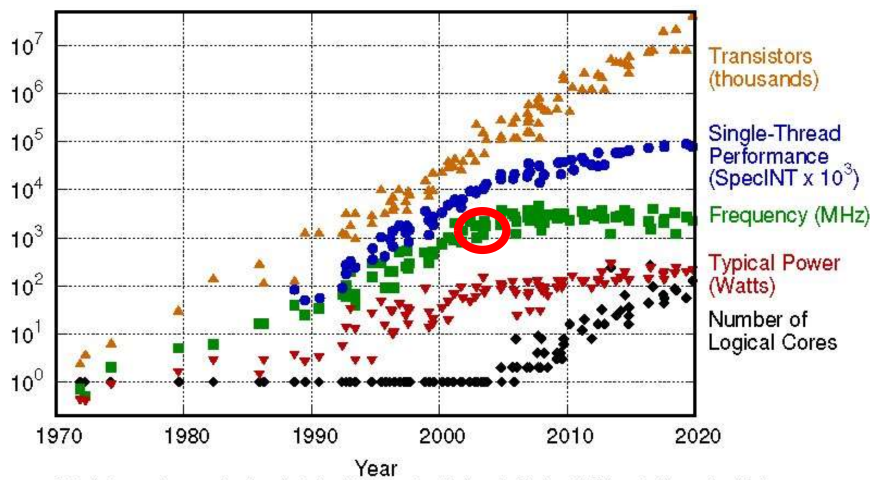
There was a clear change in the slope of the system-level performance growth trendline at the start of 2013!

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

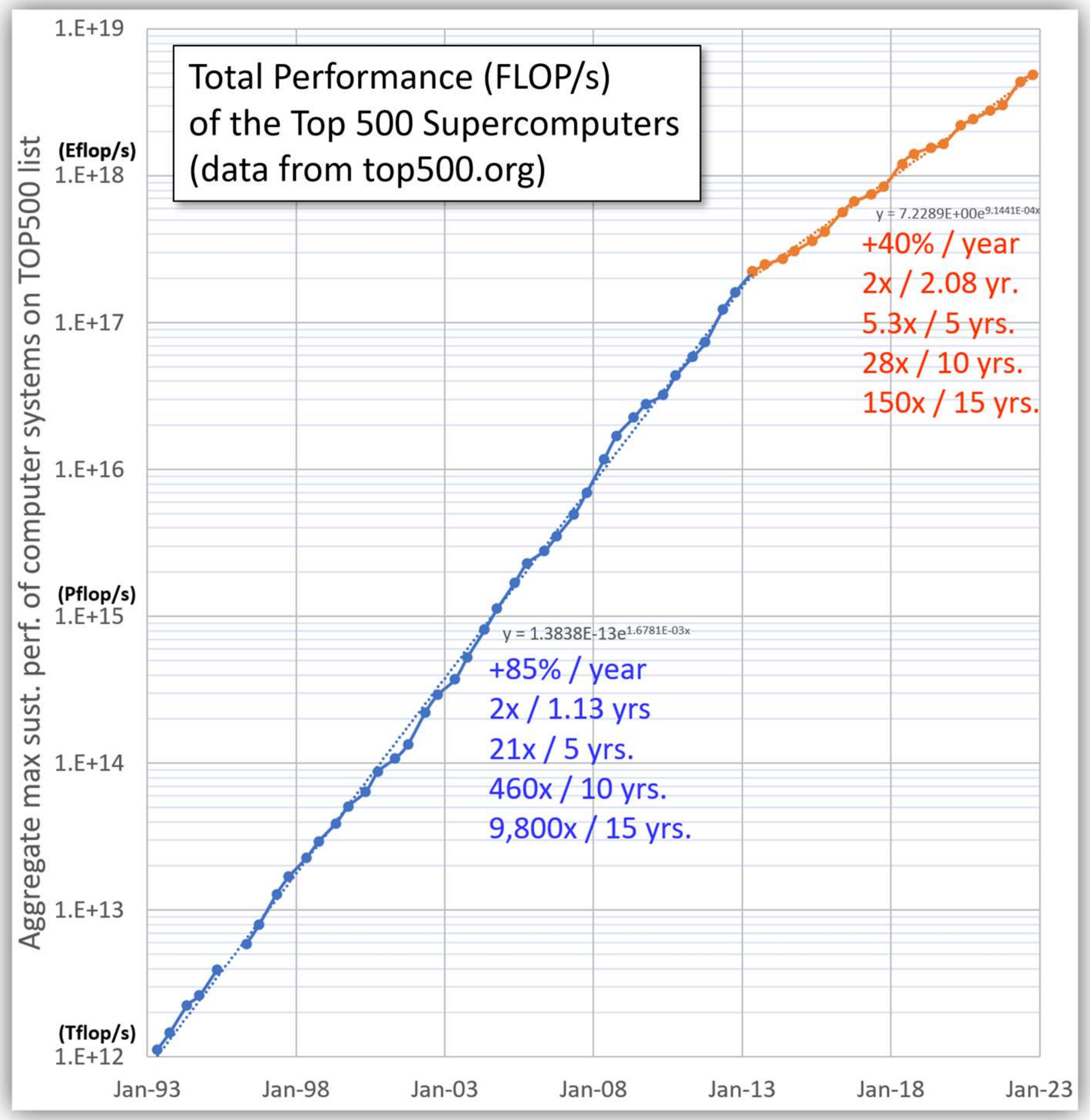
This may be attributed to a delayed system-level response to the plateauing of clock speeds that occurred in ~2005

- 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

48 Years of Microprocessor Trend Data



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2019 by K. Rupp



Total Performance (FLOP/s)
of the Top 500 Supercomputers
(data from top500.org)

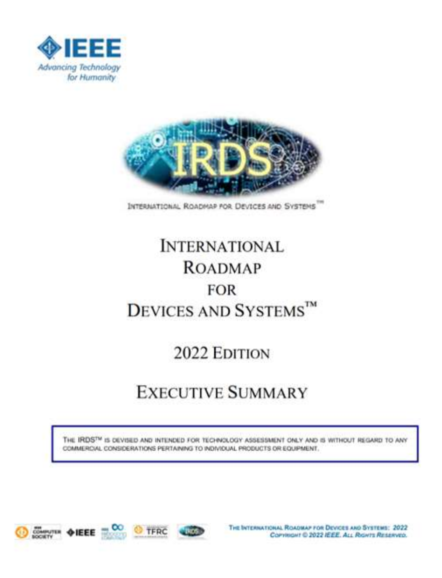
+40% / year
2x / 2.08 yr.
5.3x / 5 yrs.
28x / 10 yrs.
150x / 15 yrs.

+85% / year
2x / 1.13 yrs
21x / 5 yrs.
460x / 10 yrs.
9,800x / 15 yrs.

Looking forward now…

International Roadmap for Devices & Systems (IRDS)

irds.ieee.org



IEEE Advancing Technology for Humanity

IRDS International Roadmap for Devices and Systems™

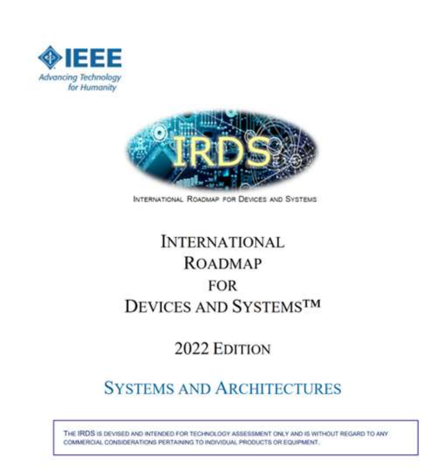
International Roadmap for Devices and Systems™

2022 Edition

Executive Summary

The IRDS™ is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

The International Roadmap for Devices and Systems: 2022
Copyright © 2022 IEEE. All Rights Reserved.



IEEE Advancing Technology for Humanity

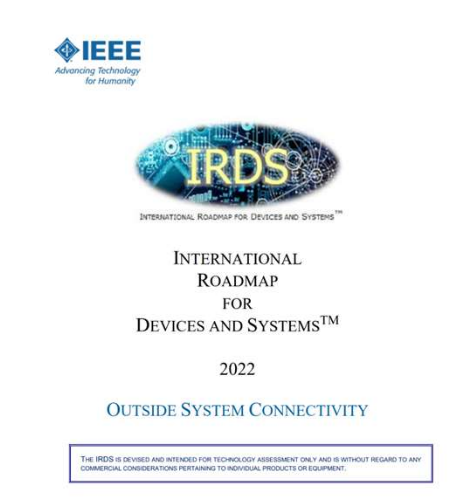
IRDS International Roadmap for Devices and Systems

International Roadmap for Devices and Systems™

2022 Edition

Systems and Architectures

The IRDS is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.



IEEE Advancing Technology for Humanity

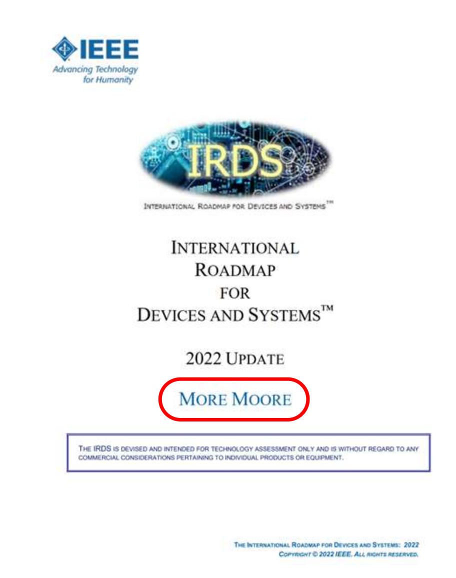
IRDS International Roadmap for Devices and Systems™

International Roadmap for Devices and Systems™

2022

Outside System Connectivity

The IRDS is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.



IEEE Advancing Technology for Humanity

IRDS International Roadmap for Devices and Systems™

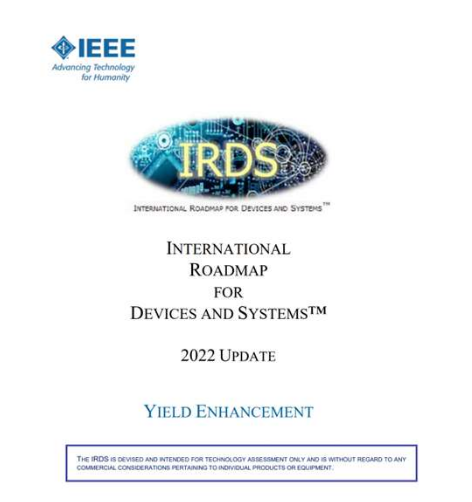
International Roadmap for Devices and Systems™

2022 Update

More Moore

The IRDS is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

The International Roadmap for Devices and Systems: 2022
Copyright © 2022 IEEE. All rights reserved.



IEEE Advancing Technology for Humanity

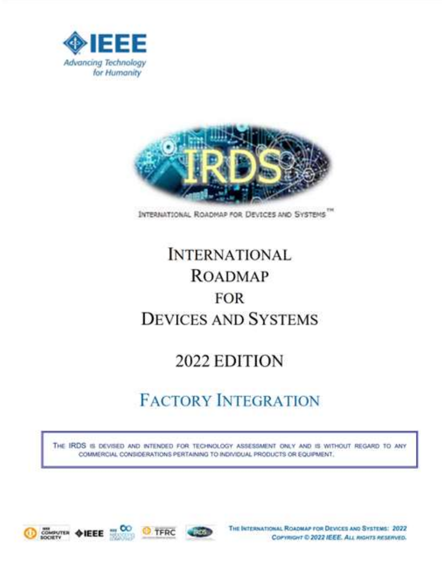
IRDS International Roadmap for Devices and Systems™

International Roadmap for Devices and Systems™

2022 Update

Yield Enhancement

The IRDS is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.



IEEE Advancing Technology for Humanity

IRDS International Roadmap for Devices and Systems™

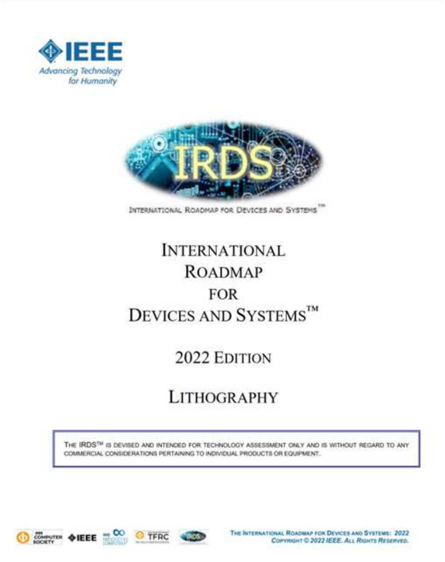
International Roadmap for Devices and Systems

2022 EDITION

Factory Integration

The IRDS is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

The International Roadmap for Devices and Systems: 2022
Copyright © 2022 IEEE. All rights reserved.



IEEE Advancing Technology for Humanity

IRDS International Roadmap for Devices and Systems™

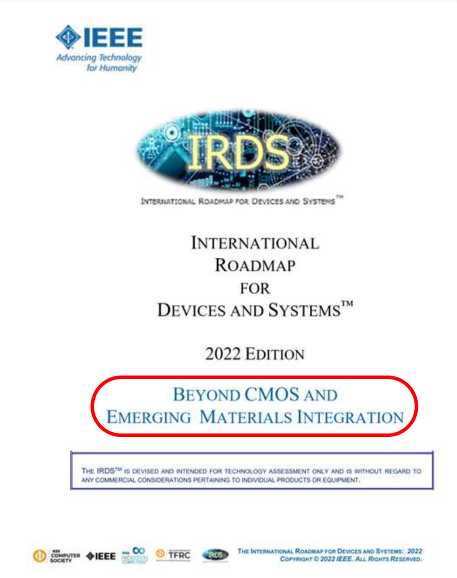
International Roadmap for Devices and Systems™

2022 Edition

Lithography

The IRDS™ is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

The International Roadmap for Devices and Systems: 2022
Copyright © 2022 IEEE. All Rights Reserved.



IEEE Advancing Technology for Humanity

IRDS International Roadmap for Devices and Systems™

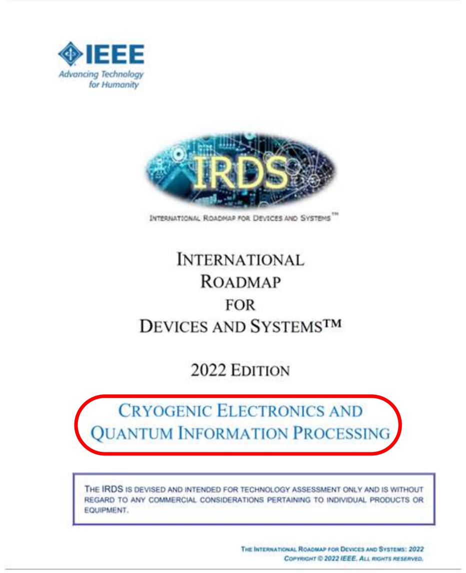
International Roadmap for Devices and Systems™

2022 Edition

Beyond CMOS and
Emerging Materials Integration

The IRDS™ is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

The International Roadmap for Devices and Systems: 2022
Copyright © 2022 IEEE. All Rights Reserved.



IEEE Advancing Technology for Humanity

IRDS International Roadmap for Devices and Systems™

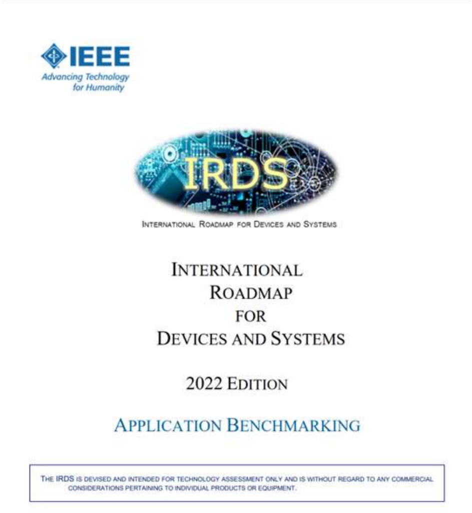
International Roadmap for Devices and Systems™

2022 Edition

Cryogenic Electronics and
Quantum Information Processing

The IRDS is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

The International Roadmap for Devices and Systems: 2022
Copyright © 2022 IEEE. All rights reserved.



IEEE Advancing Technology for Humanity

IRDS International Roadmap for Devices and Systems

International Roadmap for Devices and Systems

2022 Edition

Application Benchmarking

The IRDS is devised and intended for technology assessment only and is without regard to any commercial considerations pertaining to individual products or equipment.

Plus additional chapters and white papers…

The “**More Moore**” chapter – specifies technology node targets

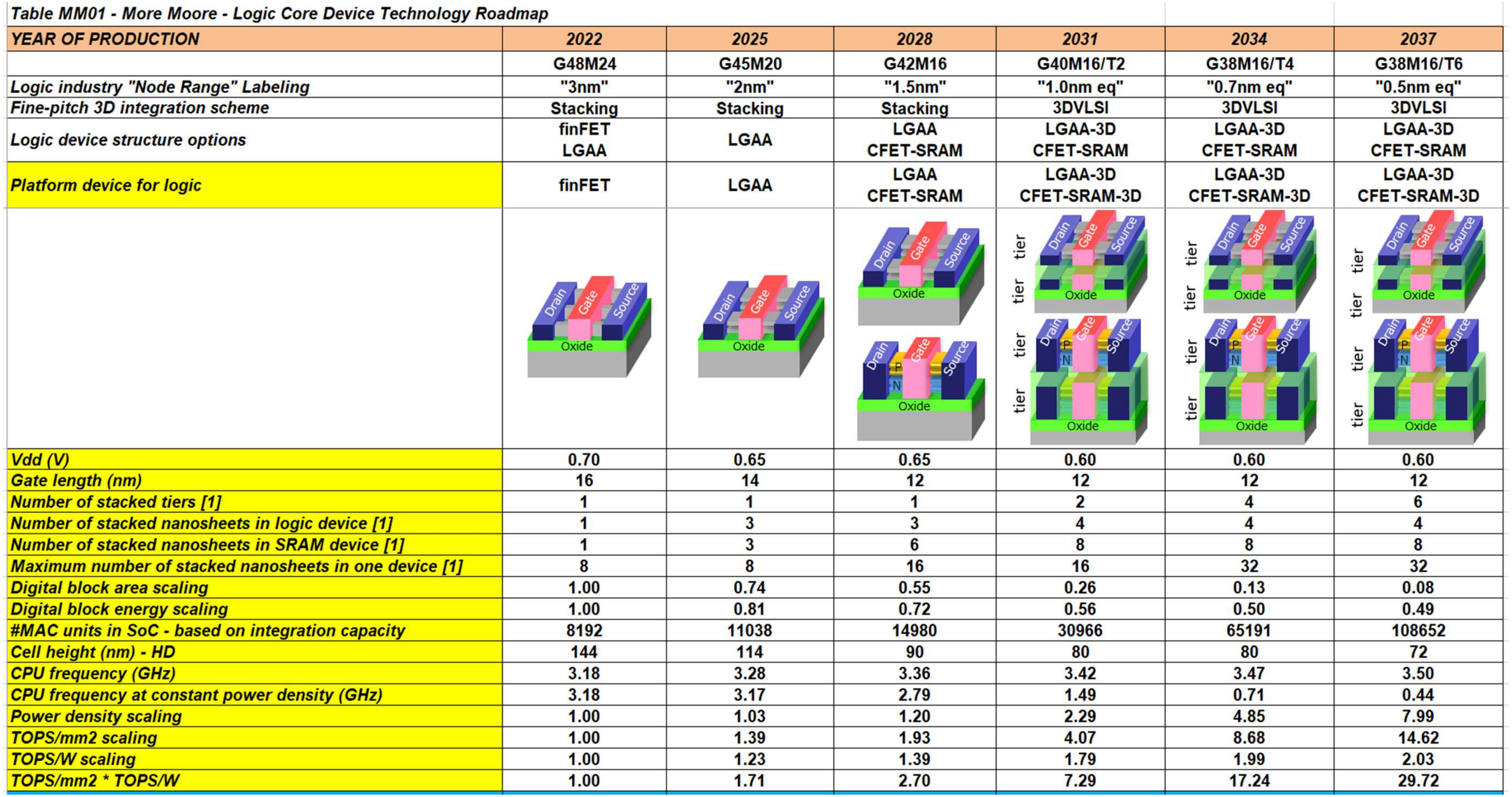
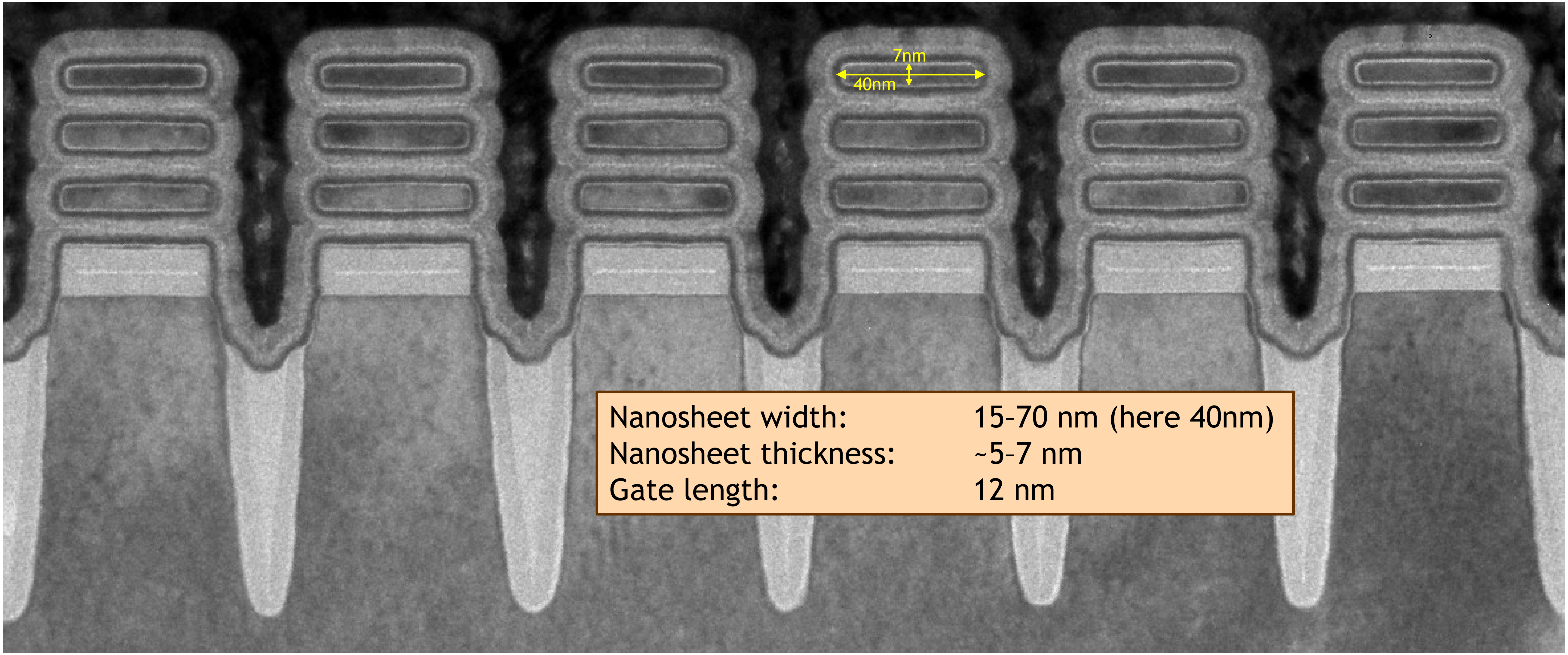


Table MM01 - More Moore - Logic Core Device Technology Roadmap

YEAR OF PRODUCTION	*2022*	*2025*	*2028*	*2031*	*2034*	*2037*
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
Platform device for logic	finFET	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D	LGAA-3D CFET-SRAM-3D
Vdd (V)	0.70	0.65	0.65	0.60	0.60	0.60
Gate length (nm)	16	14	12	12	12	12
Number of stacked tiers [1]	1	1	1	2	4	6
Number of stacked nanosheets in logic device [1]	1	3	3	4	4	4
Number of stacked nanosheets in SRAM device [1]	1	3	6	8	8	8
Maximum number of stacked nanosheets in one device [1]	8	8	16	16	32	32
Digital block area scaling	1.00	0.74	0.55	0.26	0.13	0.08
Digital block energy scaling	1.00	0.81	0.72	0.56	0.50	0.49
#MAC units in SoC - based on integration capacity	8192	11038	14980	30966	65191	108652
Cell height (nm) - HD	144	114	90	80	80	72
CPU frequency (GHz)	3.18	3.28	3.36	3.42	3.47	3.50
CPU frequency at constant power density (GHz)	3.18	3.17	2.79	1.49	0.71	0.44
Power density scaling	1.00	1.03	1.20	2.29	4.85	7.99
TOPS/mm2 scaling	1.00	1.39	1.93	4.07	8.68	14.62
TOPS/W scaling	1.00	1.23	1.39	1.79	1.99	2.03
*TOPS/mm2 * TOPS/W*	1.00	1.71	2.70	7.29	17.24	29.72

The Modern Transistor: Nanosheet Gate-All-Around (GAA) FET

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



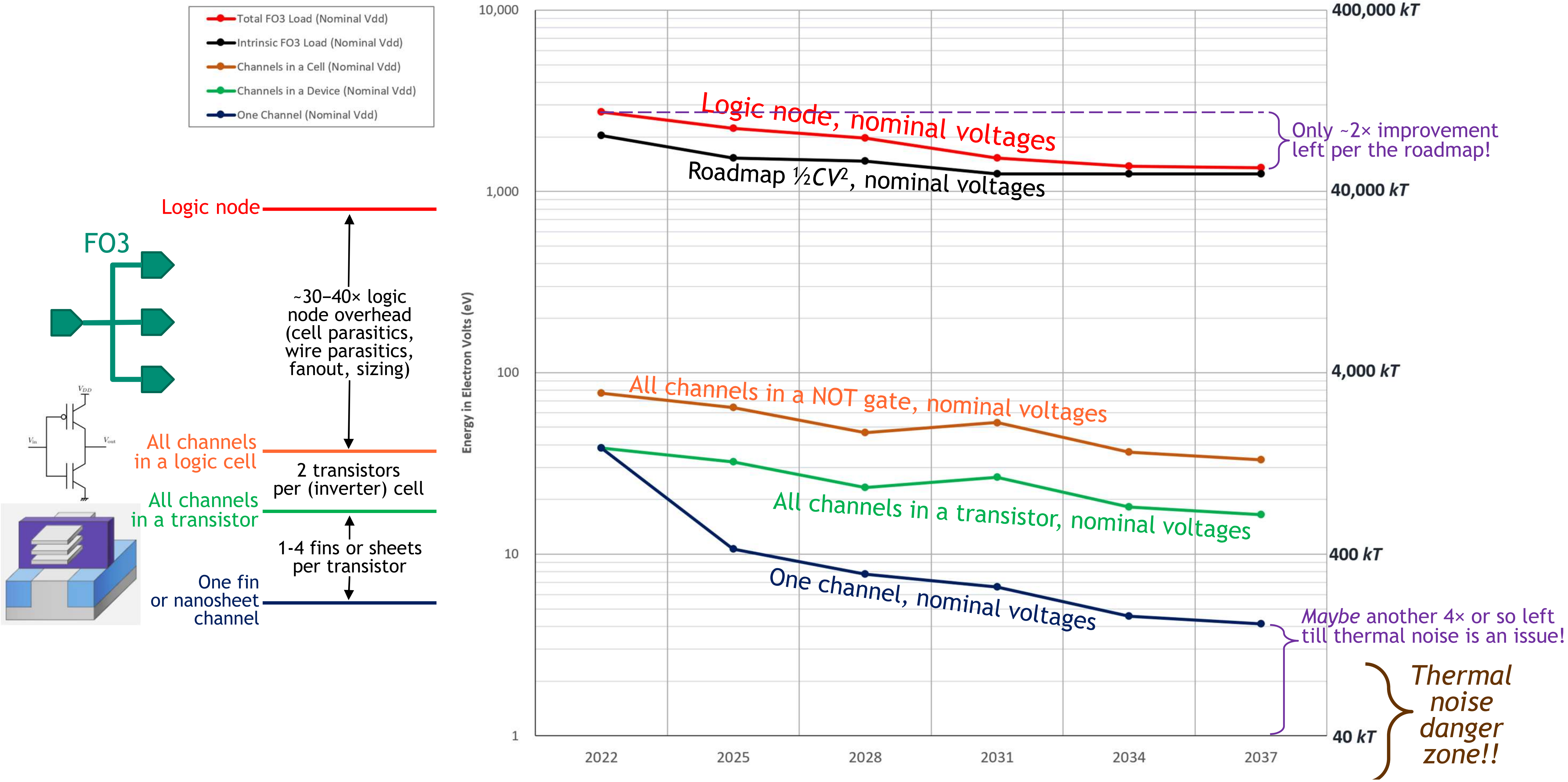
Nanosheet width:	15–70 nm (here 40nm)
Nanosheet thickness:	~5–7 nm
Gate length:	12 nm

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

(ref. temperature *T* = 290 K = 62.5 degF)

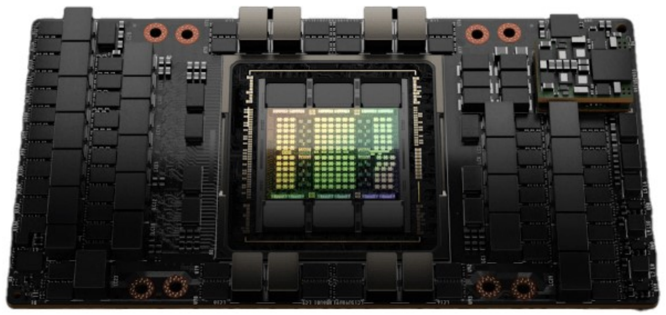


Only ~2× improvement left per the roadmap!

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

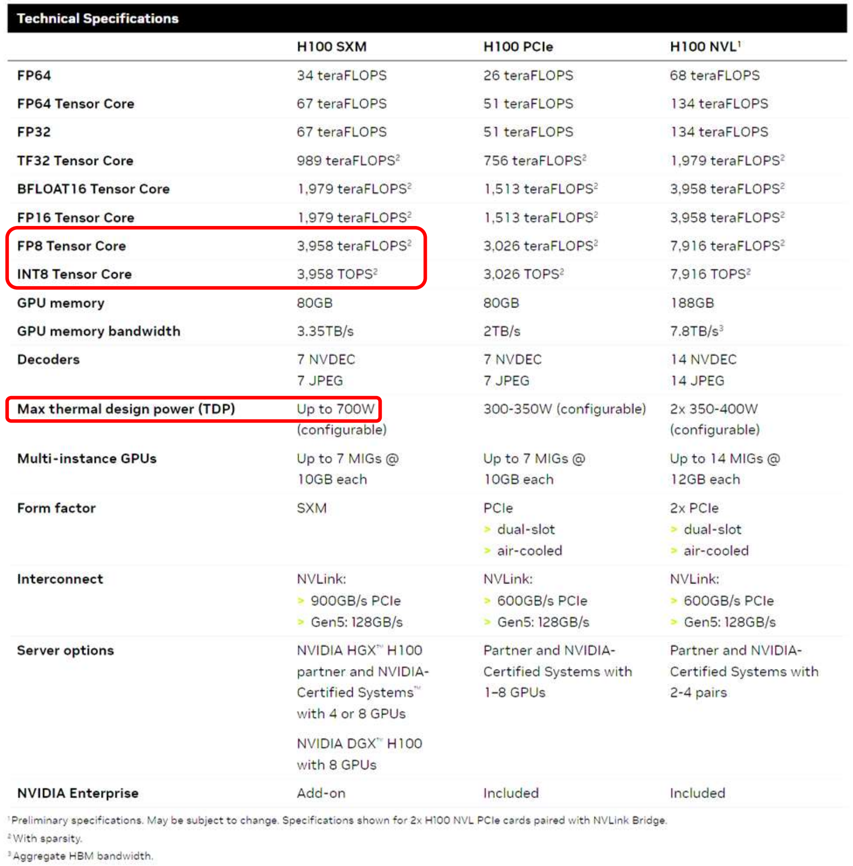
Thermal noise danger zone!!

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.)

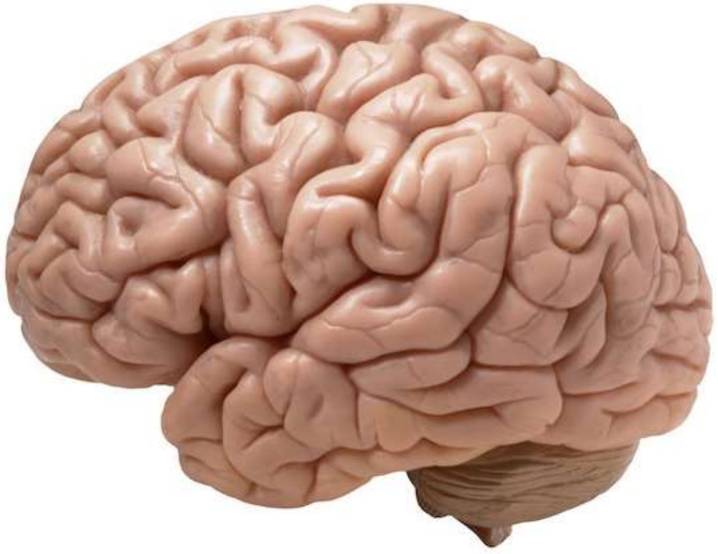


Technical Specifications	H100 SXM	H100 PCIe	H100 NVL[1]
FP64	34 teraFLOPS	26 teraFLOPS	68 teraFLOPS
FP64 Tensor Core	67 teraFLOPS	51 teraFLOPS	134 teraFLOPS
FP32	67 teraFLOPS	51 teraFLOPS	134 teraFLOPS
TF32 Tensor Core	989 teraFLOPS[2]	756 teraFLOPS[2]	1,979 teraFLOPS[2]
BFLOAT16 Tensor Core	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]	3,958 teraFLOPS[2]
FP16 Tensor Core	1,979 teraFLOPS[2]	1,513 teraFLOPS[2]	3,958 teraFLOPS[2]
FP8 Tensor Core	3,958 teraFLOPS[2]	3,026 teraFLOPS[2]	7,916 teraFLOPS[2]
INT8 Tensor Core	3,958 TOPS[2]	3,026 TOPS[2]	7,916 TOPS[2]
GPU memory	80GB	80GB	188GB
GPU memory bandwidth	3.35TB/s	2TB/s	7.8TB/s[3]
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG	14 NVDEC 14 JPEG
Max thermal design power (TDP)	Up to 700W (configurable)	300-350W (configurable)	2x 350-400W (configurable)
Multi-instance GPUs	Up to 7 MIGs @ 10GB each	Up to 7 MIGs @ 10GB each	Up to 14 MIGs @ 12GB each
Form factor	SXM	PCIe > dual-slot > air-cooled	2x PCIe > dual-slot > air-cooled
Interconnect	NVLink: > 900GB/s PCIe > Gen5: 128GB/s	NVLink: > 600GB/s PCIe > Gen5: 128GB/s	NVLink: > 600GB/s PCIe > Gen5: 128GB/s
Server options	NVIDIA HGX™ H100 partner and NVIDIA-Certified Systems™ with 4 or 8 GPUs NVIDIA DGX™ H100 with 8 GPUs	Partner and NVIDIA-Certified Systems with 1–8 GPUs	Partner and NVIDIA-Certified Systems with 2-4 pairs
NVIDIA Enterprise	Add-on	Included	Included

[1] Preliminary specifications. May be subject to change. Specifications shown for 2x H100 NVL PCIe cards paired with NVLink Bridge.
[2] With sparsity.
[3] Aggregate HBM bandwidth.

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! 😲

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!

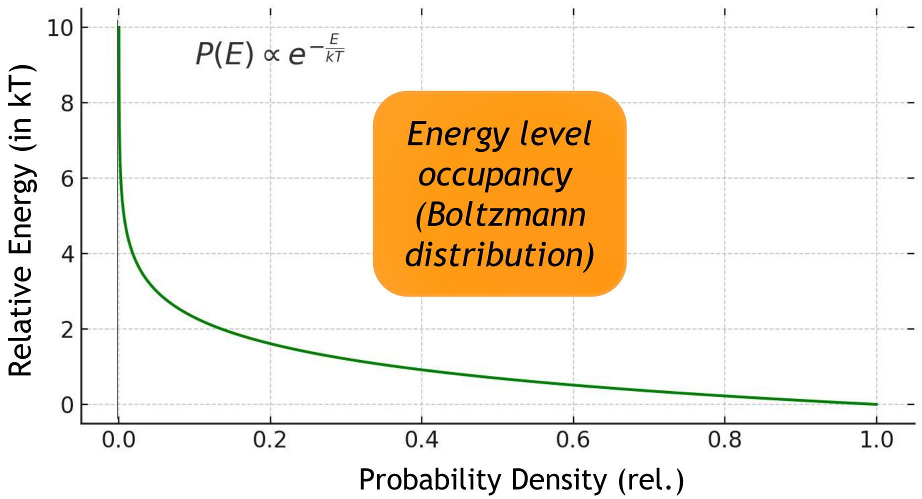
Coincidence? 🤔 Or not?

Fundamental Physics behind the Limits of CMOS

Thermal fluctuations and the Boltzmann distribution

- Discovery of thermal fluctuations
  - Robert Brown (1827): Empirical observation of erratic motion in pollen grains, known as "Brownian motion."
  - 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:*

$$f(E) = Ae^{-E/kT}$$



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:

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

$$r_{\mathrm{on/off}} \overset{\mathrm{def}}{=} \frac{g_{\mathrm{ds}}(V_{\mathrm{gs}} = V_{\mathrm{dd}})}{g_{\mathrm{ds}}(V_{\mathrm{gs}} = 0)}$$

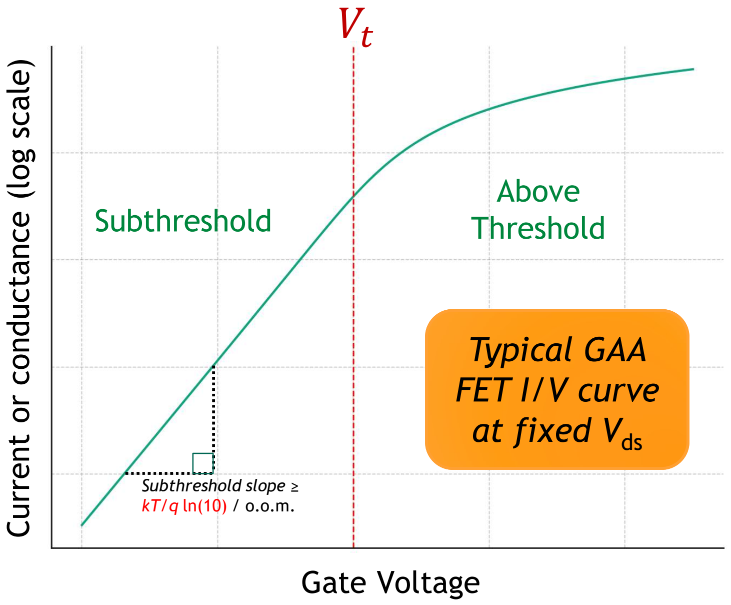
$$r_{\mathrm{on/off}} \leq \exp\left(\frac{V_{\mathrm{dd}}}{kT/q}\right)$$

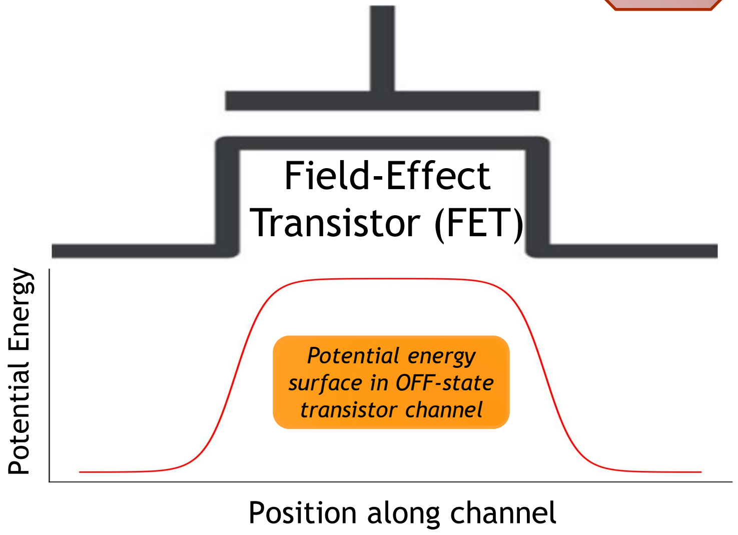
$$V_{\mathrm{dd}} \geq \frac{kT}{q}\ln(r_{\mathrm{on/off}})$$

$$E_{\mathrm{sw}} \geq kT\ln(r_{\mathrm{on/off}})$$

(per electron in channel!)

~30 kT /channel in 2028

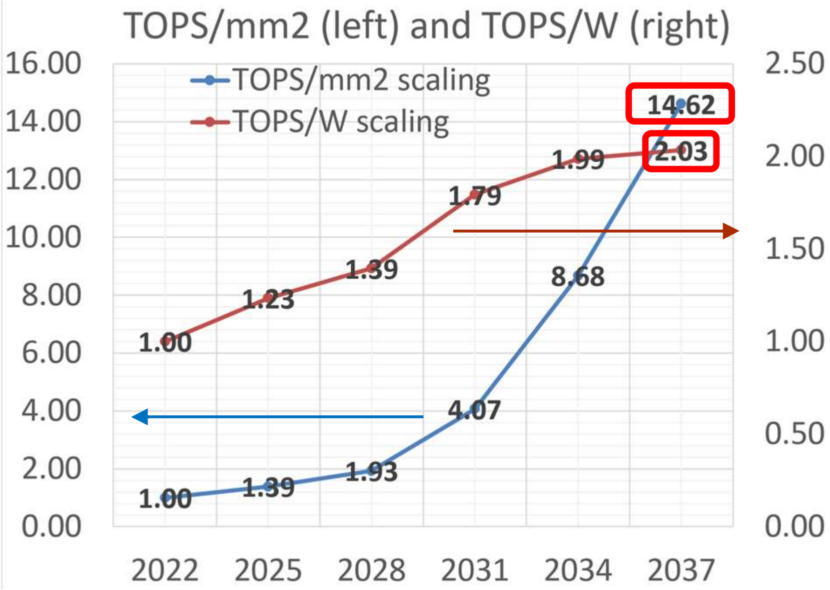



Field-Effect Transistor (FET)
Potential energy surface in OFF-state transistor channel
Potential Energy
Position along channel

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…

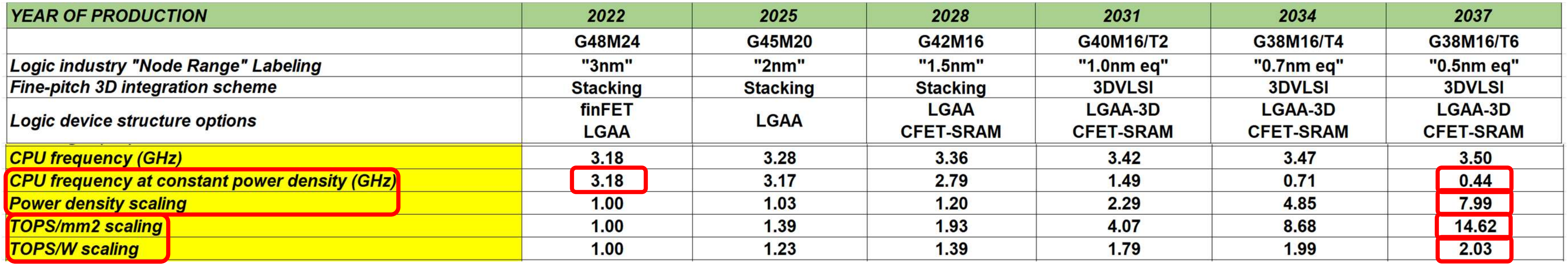
- Note that on the roadmap, efficiency is only improving **2×** by 2037, and they project that for throughput density to increase by the maximum of ~14.6×, power dissipation density would have to increase by **~8×**!
  - Imagine trying to cool a GPU chip of fixed area that now dissipates 5,600 W instead of 700 W!
- Or, if what we want is to keep power density constant, processor clock speeds would have to fall ~7.2×—*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×**).



(Source: IRDS '22 More Moore chapter)

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!

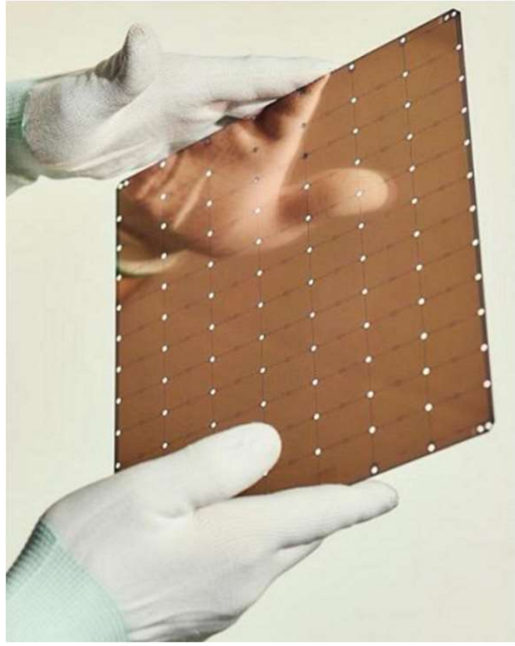


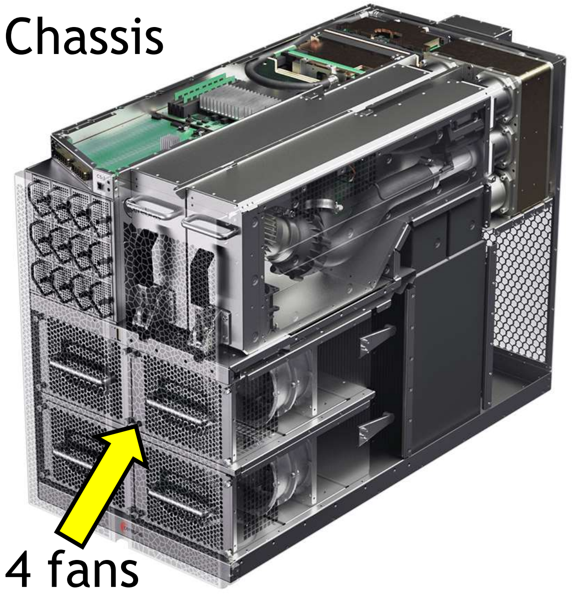
YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
	G48M24	G45M20	G42M16	G40M16/T2	G38M16/T4	G38M16/T6
Logic industry "Node Range" Labeling	"3nm"	"2nm"	"1.5nm"	"1.0nm eq"	"0.7nm eq"	"0.5nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	finFET LGAA	LGAA	LGAA CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM	LGAA-3D CFET-SRAM
CPU frequency (GHz)	3.18	3.28	3.36	3.42	3.47	3.50
CPU frequency at constant power density (GHz)	3.18	3.17	2.79	1.49	0.71	0.44
Power density scaling	1.00	1.03	1.20	2.29	4.85	7.99
TOPS/mm2 scaling	1.00	1.39	1.93	4.07	8.68	14.62
TOPS/W scaling	1.00	1.23	1.39	1.79	1.99	2.03

Cooling system designs are already starting to get insane as it is…

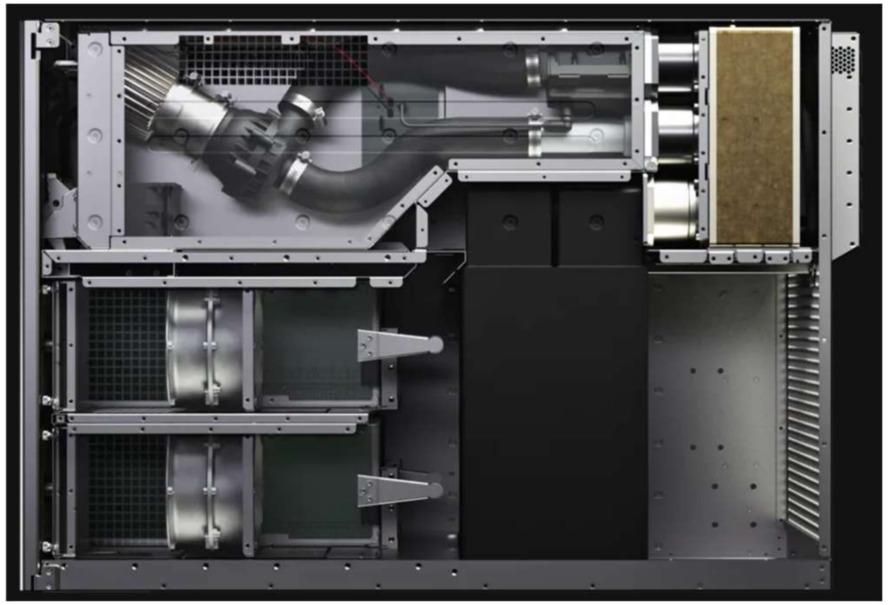
E.g., Cerebras WSE-2 is the largest, highest-performing single AI chip today… BUT it uses up to **23 kW**!

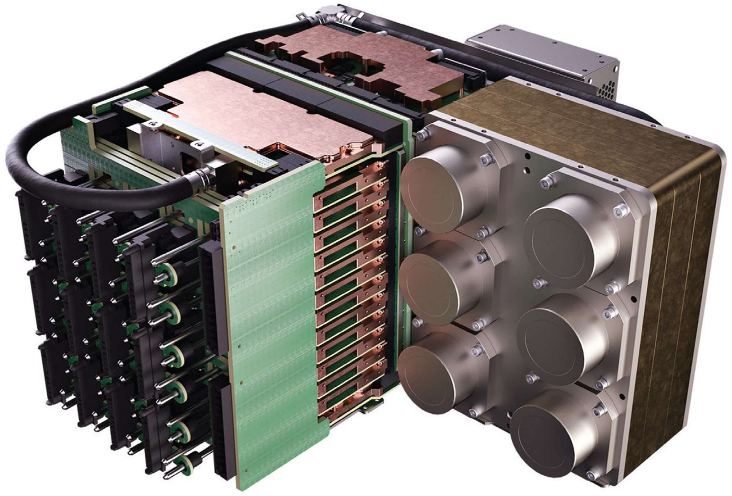
- And just look at what-all that requires in terms of cooling hardware <u>already</u>… *(How would you boost this another 8x* → **184 kW***?)*

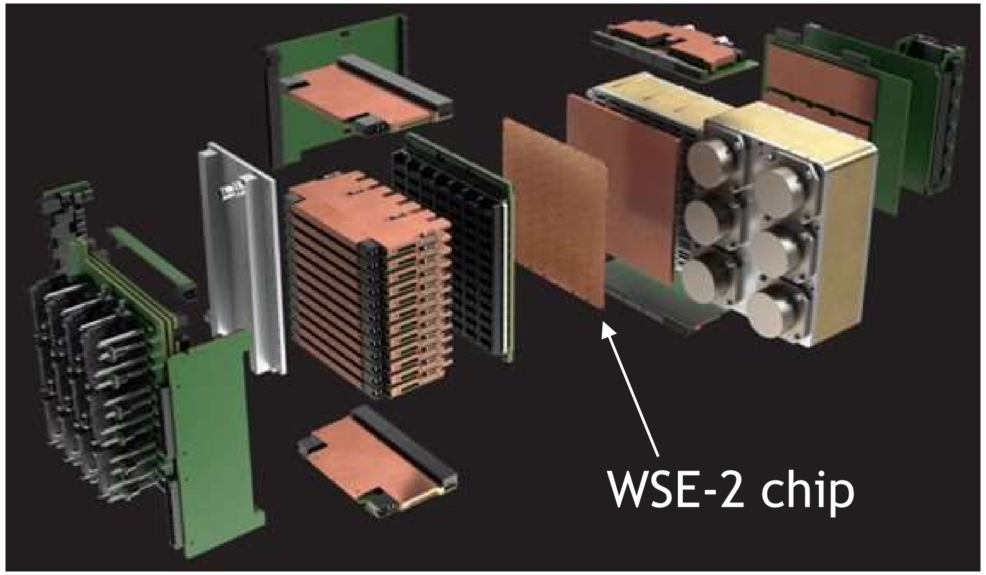


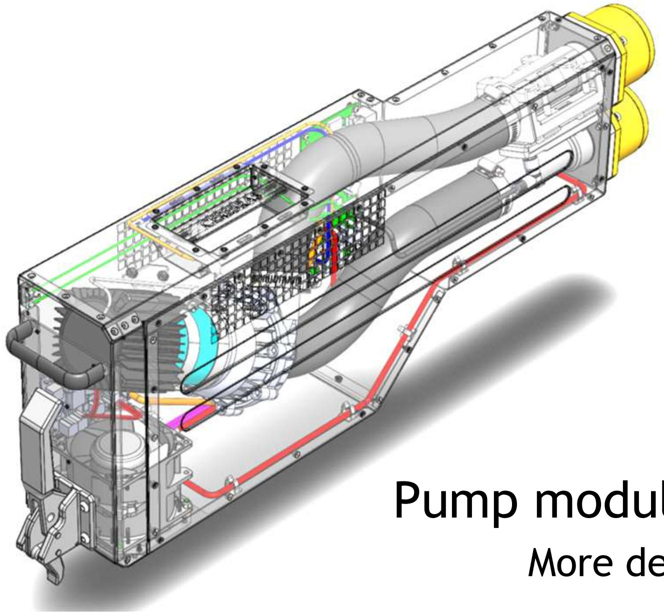

Chassis

4 fans

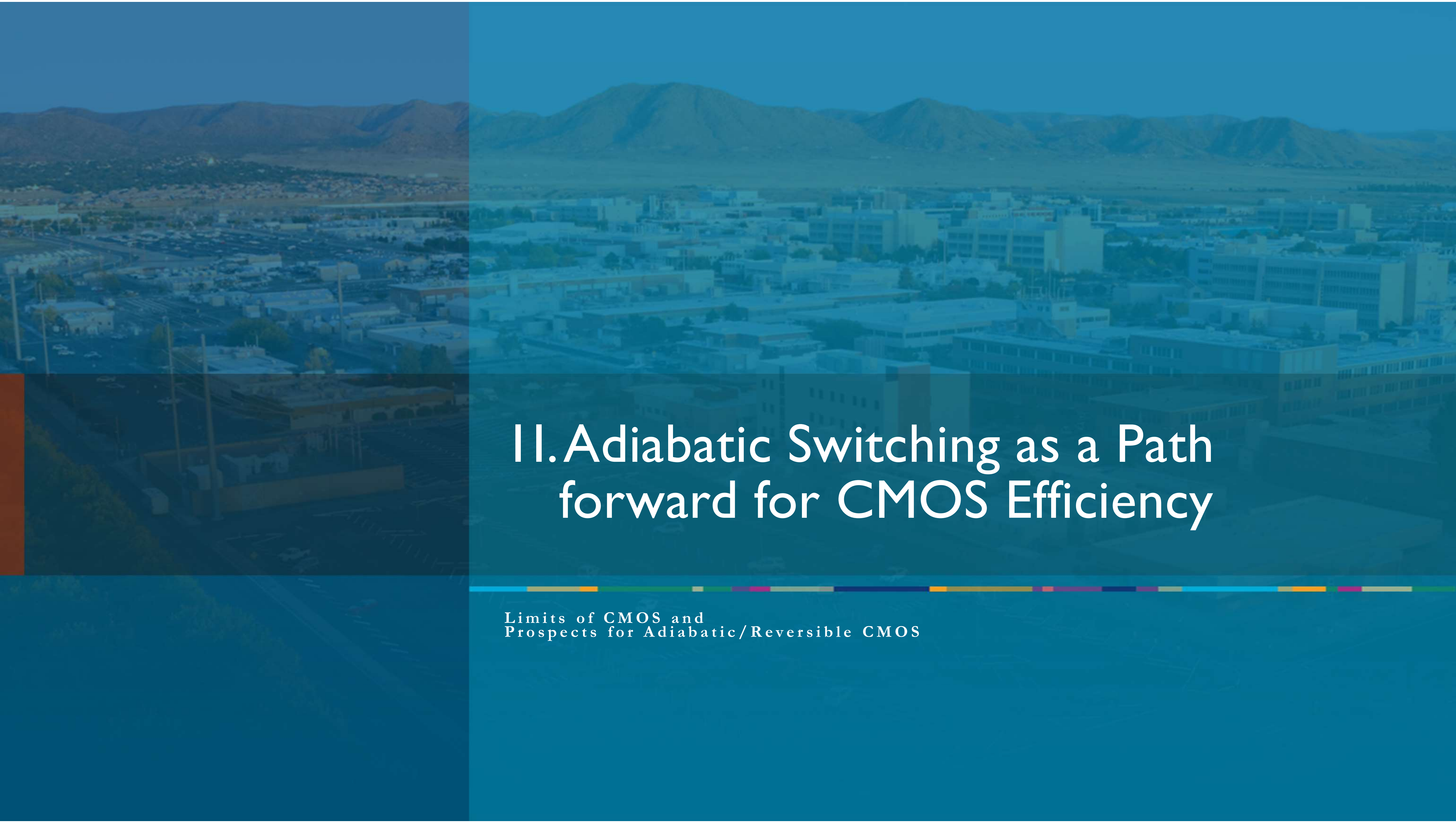



"Engine Block"




Pump module

More details →



II. Adiabatic Switching as a Path forward for CMOS Efficiency

Limits of CMOS and
Prospects for Adiabatic/Reversible CMOS

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 ½CV^2 in each (sudden) switching event…

- Consequence of $Q = CV$ charge delivered from voltage V → later returned to 0V.

In *adiabatic* CMOS, we instead deliver charge in a gradual, steady flow…

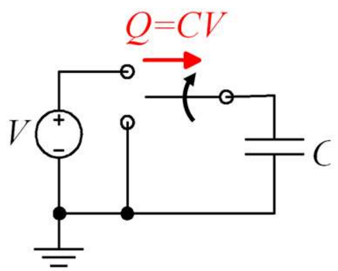
- We can think of the source as being *constant current* instead of *constant voltage*.
- Can approximate constant current source with a ~linear *voltage ramp* over time t.
- Because the charge transfer is more gradual, the voltage drop over the charging path is smaller, and so the energy dissipated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

We basically can make the energy dissipation <u>as small as we want</u>.

- Down to a lower limit set by leakage.

Conventional charging:

- Constant *voltage* source

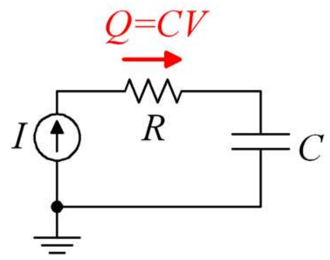


- Energy dissipated:

$$E_{\mathrm{diss}}^{\mathrm{conv}} = \frac{1}{2}CV^2$$

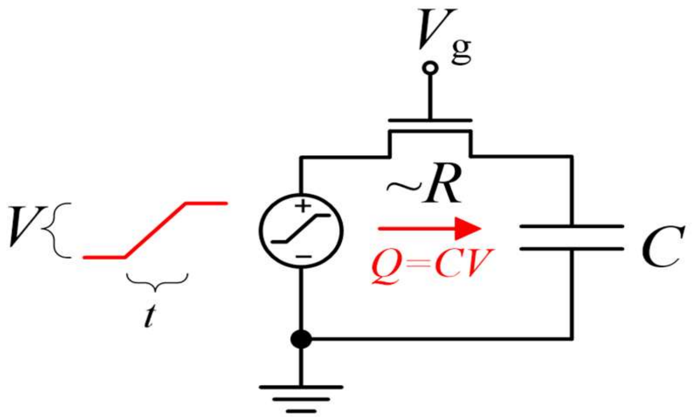
Ideal *adiabatic* charging:

- Constant *current* source



- Energy dissipated:

$$E_{\mathrm{diss}}^{\mathrm{adia}} = I^2Rt = \frac{Q^2R}{t} = CV^2\frac{RC}{t}$$



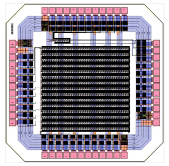
$$t \gg RC \Rightarrow E_{\mathrm{diss}} \to CV^2\frac{RC}{t}$$

$$t \ll RC \Rightarrow E_{\mathrm{diss}} \to \tfrac{1}{2}CV^2$$

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

2LAL test chip taped out at Sandia, Aug. '20



To approach ideal reversible computing in CMOS…

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

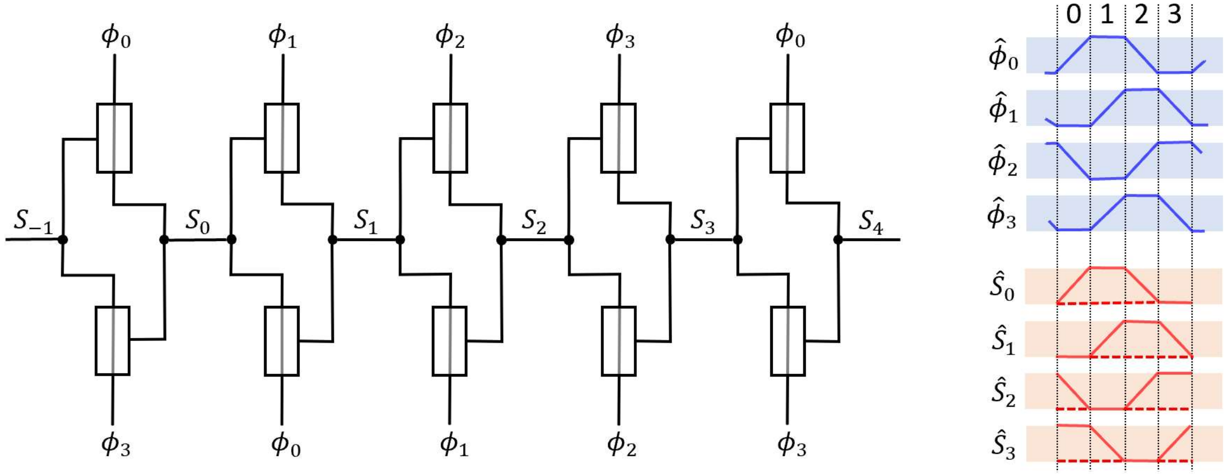
- Diodes in charging path, "sparking," "squelching,"
  - Eliminated by "**truly, fully adiabatic**" design. (*E.g.*, CRL, 2LAL).
    - Can suffice to get down to a few aJ (10s of eV) even *before* voltage optimization!
- Voltage level mismatches that dynamically arise on floating nodes before reconnection.
  - Eliminated by static, "**perfectly adiabatic**" design. (*E.g.*, S2LAL).

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

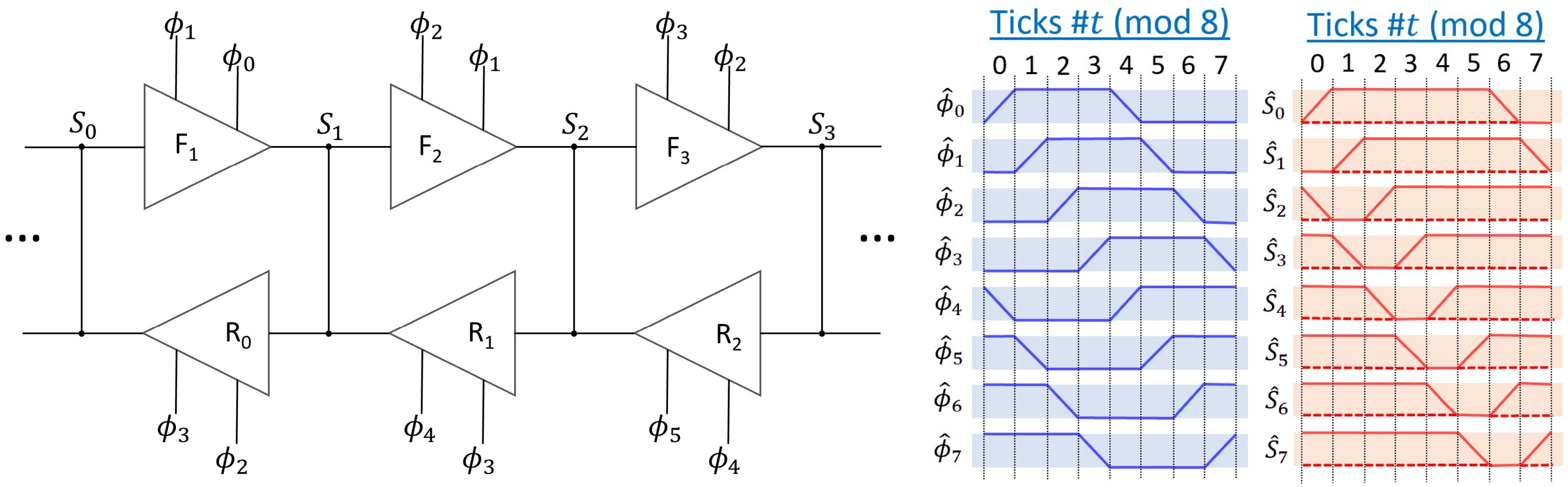
- Subthreshold channel currents.
  - Ultra-low-*T* (*e.g.* 4K) operation helps with this.
- Tunneling through gate oxide.
  - *E.g.*, use thicker gate oxides.

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

Shift Register Structure and Timing in 2LAL



Shift Register Structure and Timing in S2LAL



(arxiv:2009.00448)

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

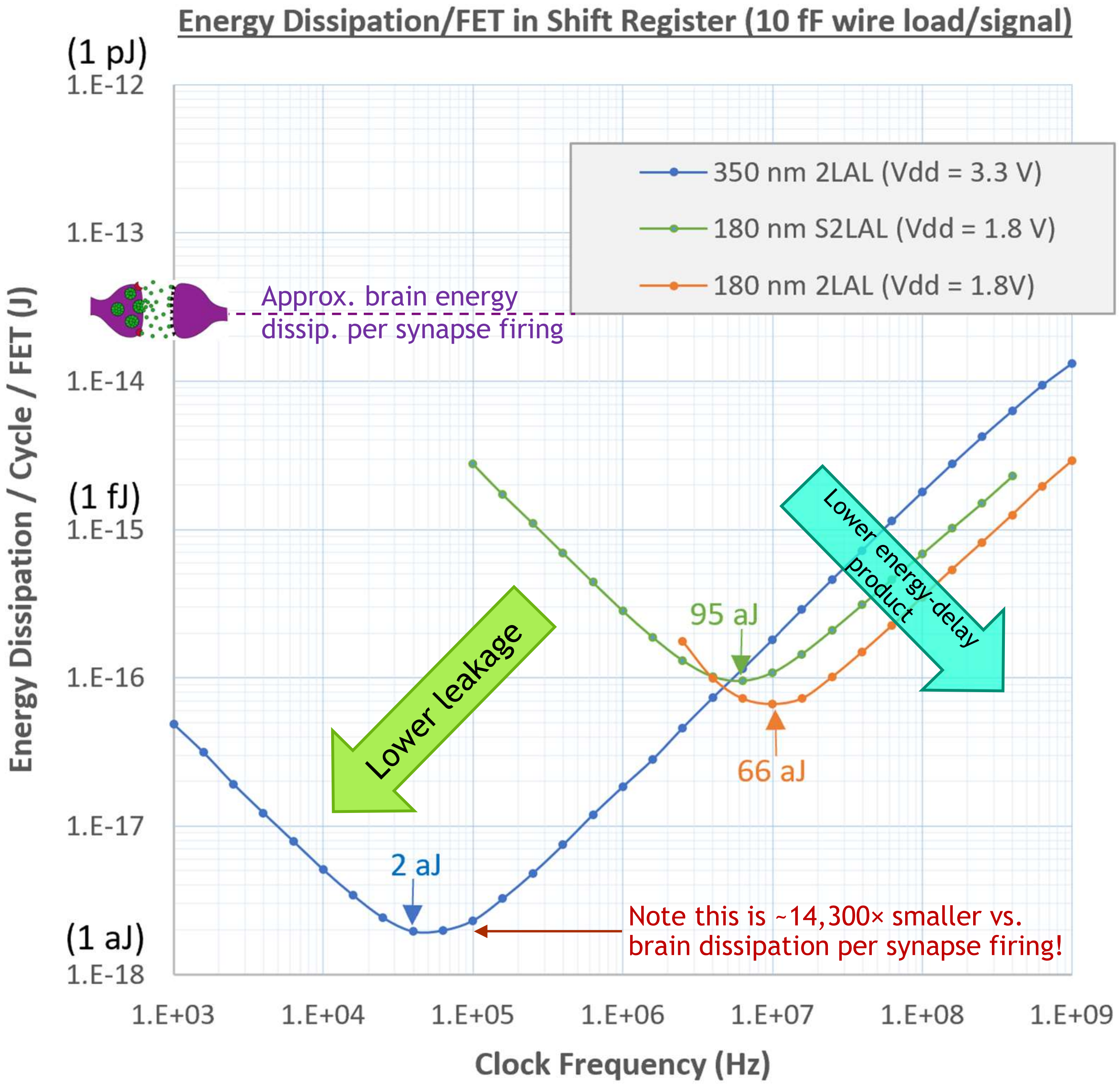
Plotted energy dissipation per-transistor in shift registers at 50% activity factor (alternating 0/1)

- 2LAL (**blue**, **orange** curves)
- S2LAL (**green** curve)

In all of these Cadence/Spectre simulations,

- We assumed a 10 fF parasitic wiring load 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}.



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

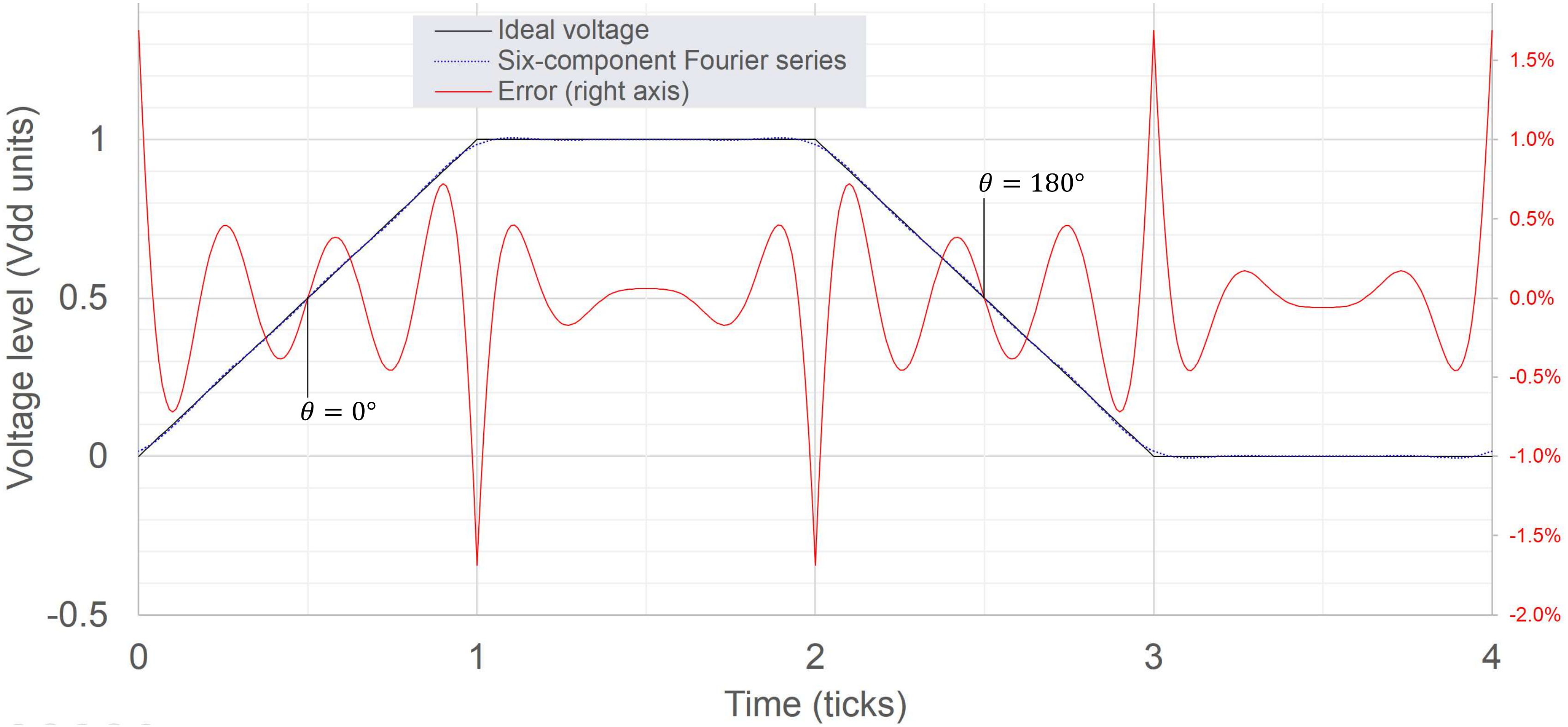
- $\therefore$ Spectrum includes only odd-numbered harmonics $f, 3f, 5f, \ldots$

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

- Maximum error with $11f$ frequency cutoff is $< 1.7\%$ of V_{dd}

Cf. S. G. Younis and T. F. Knight, "Non-dissipative rail drivers for adiabatic circuits," in Proc. Sixteenth Conf. on Advanced Research in VLSI, IEEE, Mar. 1995, pp. 404-414. doi:10.1109/ARVLSI.1995.515635

$$v_{f6}(t) = V_{\mathrm{DD}}\left[\frac{1}{2}+\frac{4\sqrt{2}}{\pi^2}\begin{pmatrix} \sin(\omega t) \\ +\dfrac{\sin(3\omega t)}{3^2} \\ -\dfrac{\sin(5\omega t)}{5^2} \\ -\dfrac{\sin(7\omega t)}{7^2} \\ +\dfrac{\sin(9\omega t)}{9^2} \\ +\dfrac{\sin(11\omega t)}{11^2} \end{pmatrix}\right]$$



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 ~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

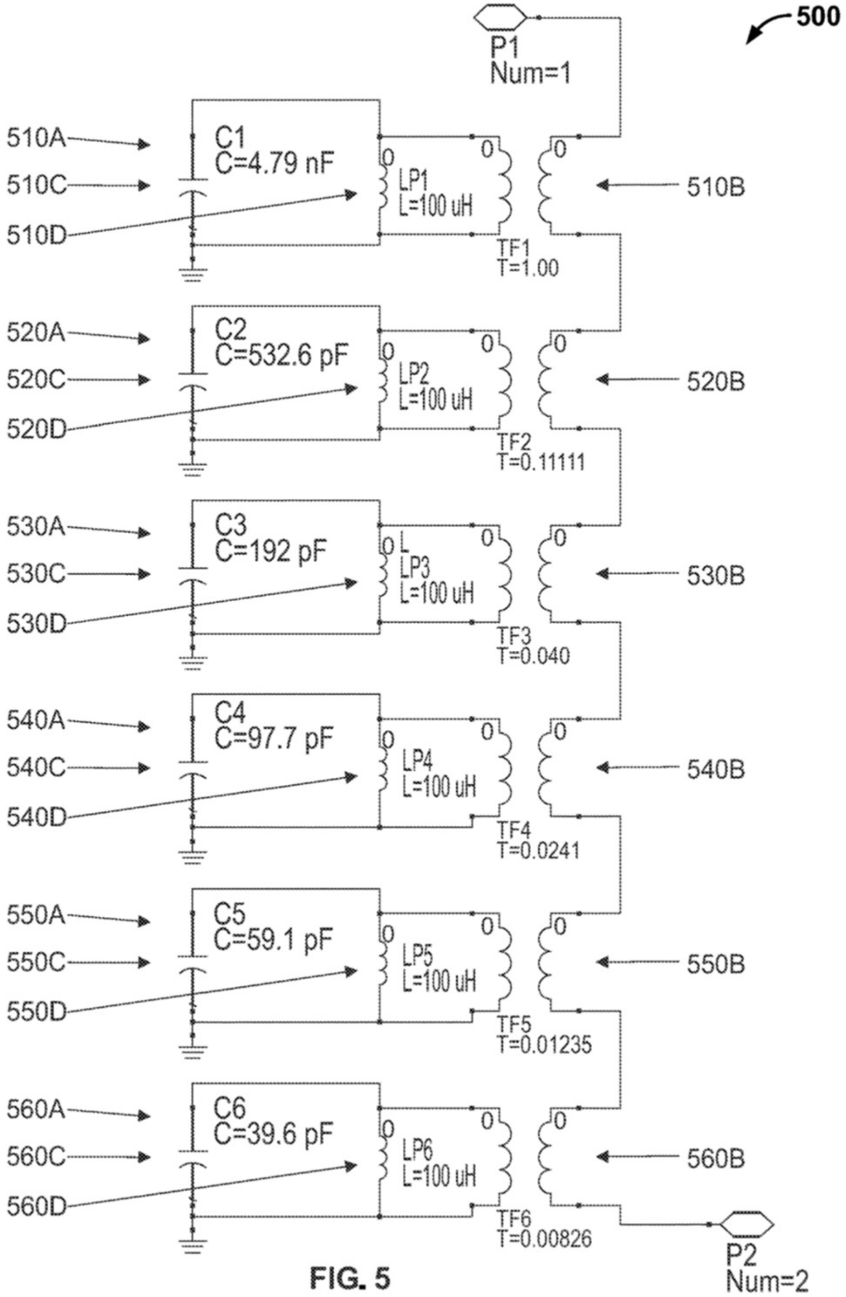
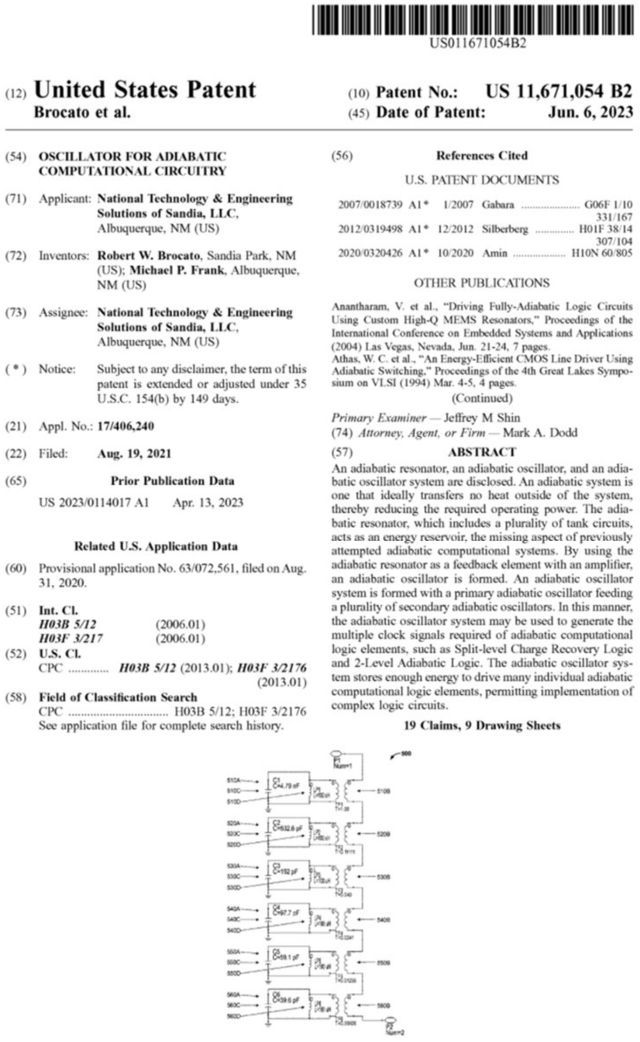


FIG. 5



US011671054B2

(12) **United States Patent**
Brocato et al.

(10) **Patent No.: US 11,671,054 B2**
(45) **Date of Patent: Jun. 6, 2023**

(54) **OSCILLATOR FOR ADIABATIC COMPUTATIONAL CIRCUITRY**

(71) Applicant: **National Technology & Engineering Solutions of Sandia, LLC**, Albuquerque, NM (US)

(72) Inventors: **Robert W. Brocato**, Sandia Park, NM (US); **Michael P. Frank**, Albuquerque, NM (US)

(73) Assignee: **National Technology & Engineering Solutions of Sandia, LLC**, Albuquerque, NM (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 149 days.

(21) Appl. No.: **17/406,240**

(22) Filed: **Aug. 19, 2021**

(65) **Prior Publication Data**

US 2023/0114017 A1 Apr. 13, 2023

Related U.S. Application Data

(60) Provisional application No. 63/072,561, filed on Aug. 31, 2020.

(51) **Int. Cl.**
H03B 5/12 (2006.01)
H03F 3/217 (2006.01)

(52) **U.S. Cl.**
CPC ***H03B 5/12*** (2013.01); ***H03F 3/2176*** (2013.01)

(58) **Field of Classification Search**
CPC H03B 5/12; H03F 3/2176
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

2007/0018739	A1 *	1/2007	Gabara	G06F 1/10 331/167
2012/0319498	A1 *	12/2012	Silberberg	H01F 38/14 307/104
2020/0320426	A1 *	10/2020	Amin	H10N 60/805

OTHER PUBLICATIONS

Anantharam, V. et al., "Driving Fully-Adiabatic Logic Circuits Using Custom High-Q MEMS Resonators," Proceedings of the International Conference on Embedded Systems and Applications (2004) Las Vegas, Nevada, Jun. 21-24, 7 pages.
Athas, W. C. et al., "An Energy-Efficient CMOS Line Driver Using Adiabatic Switching," Proceedings of the 4th Great Lakes Symposium on VLSI (1994) Mar. 4-5, 4 pages.

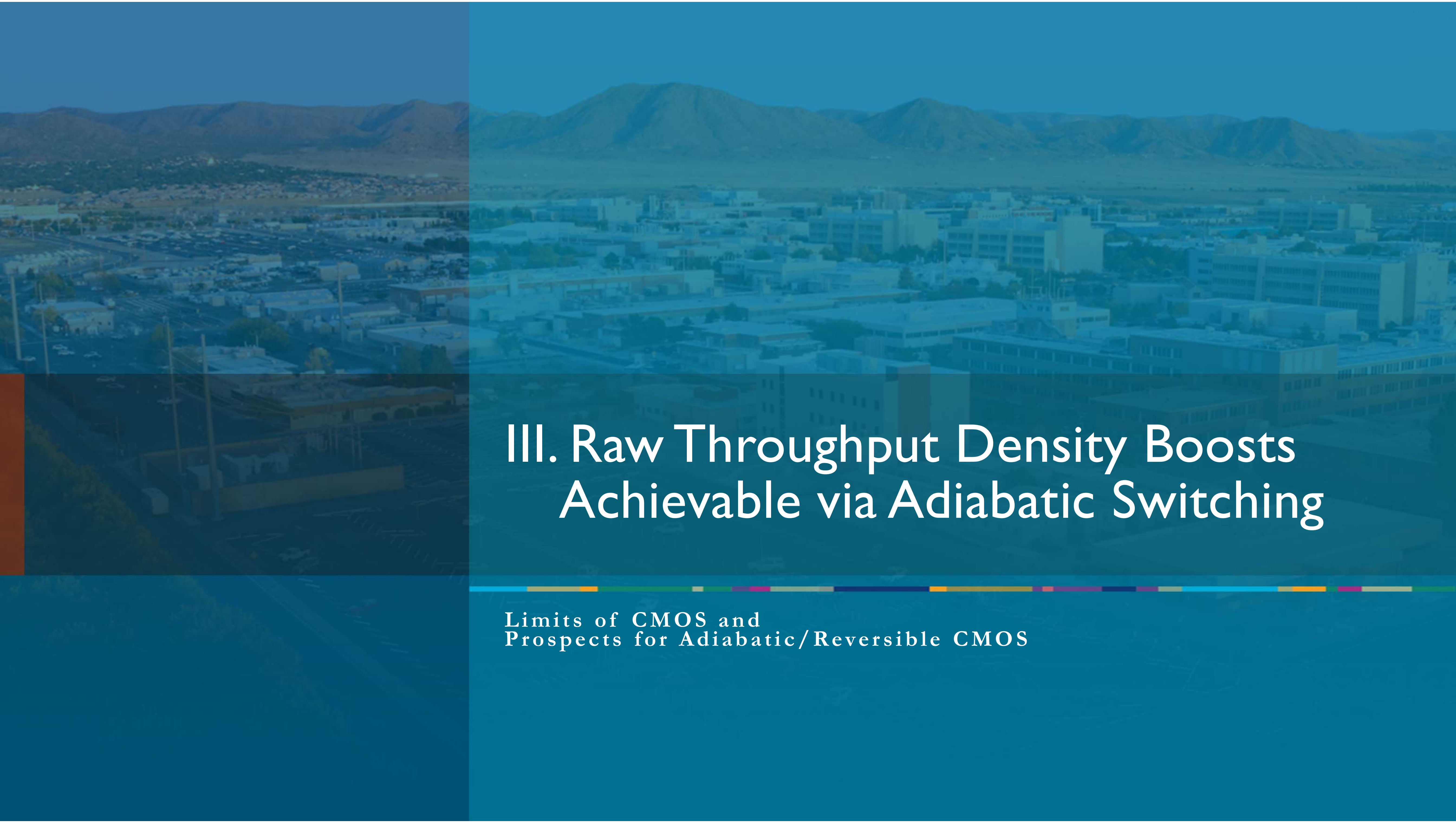
(Continued)

Primary Examiner — Jeffrey M Shin
(74) *Attorney, Agent, or Firm* — Mark A. Dodd

(57) **ABSTRACT**

An adiabatic resonator, an adiabatic oscillator, and an adiabatic oscillator system are disclosed. An adiabatic system is one that ideally transfers no heat outside of the system, thereby reducing the required operating power. The adiabatic resonator, which includes a plurality of tank circuits, acts as an energy reservoir, the missing aspect of previously attempted adiabatic computational systems. By using the adiabatic resonator as a feedback element with an amplifier, an adiabatic oscillator is formed. An adiabatic oscillator system is formed with a primary adiabatic oscillator feeding a plurality of secondary adiabatic oscillators. In this manner, the adiabatic oscillator system may be used to generate the multiple clock signals required of adiabatic computational logic elements, such as Split-level Charge Recovery Logic and 2-Level Adiabatic Logic. The adiabatic oscillator system stores enough energy to drive many individual adiabatic computational logic elements, permitting implementation of complex logic circuits.

19 Claims, 9 Drawing Sheets



III. Raw Throughput Density Boosts Achievable via Adiabatic Switching

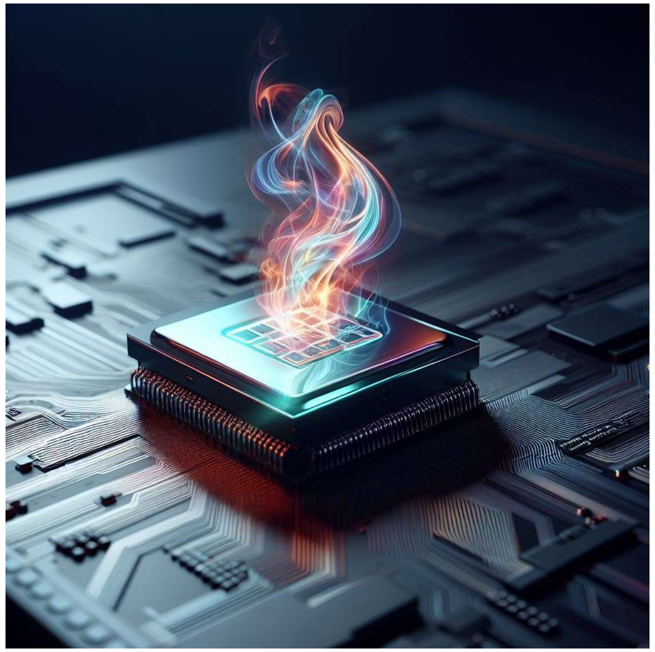
Limits of CMOS and
Prospects for Adiabatic/Reversible CMOS

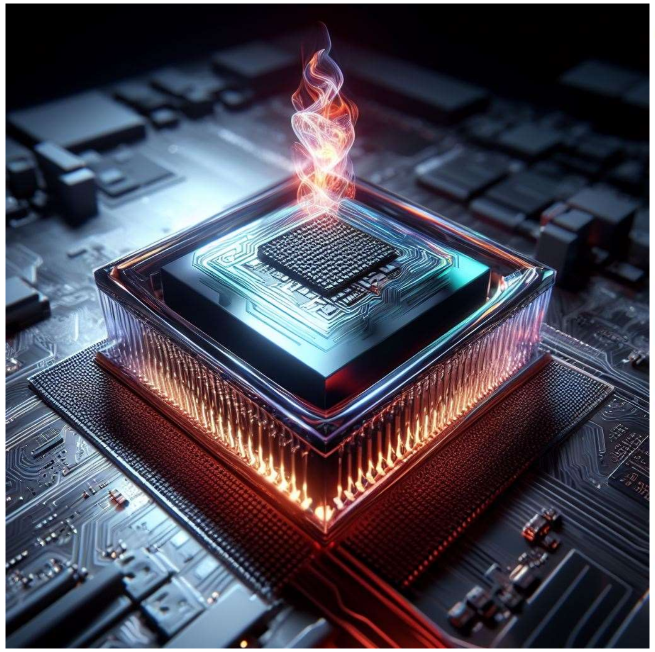
Analysis of Throughput Density Boost from Adiabatic Switching

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 → 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.)





Note: All results shown on this & the next few slides are for the IRDS's "HD" (high-density) design scenario

Standard-Voltage Conventional Switching

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

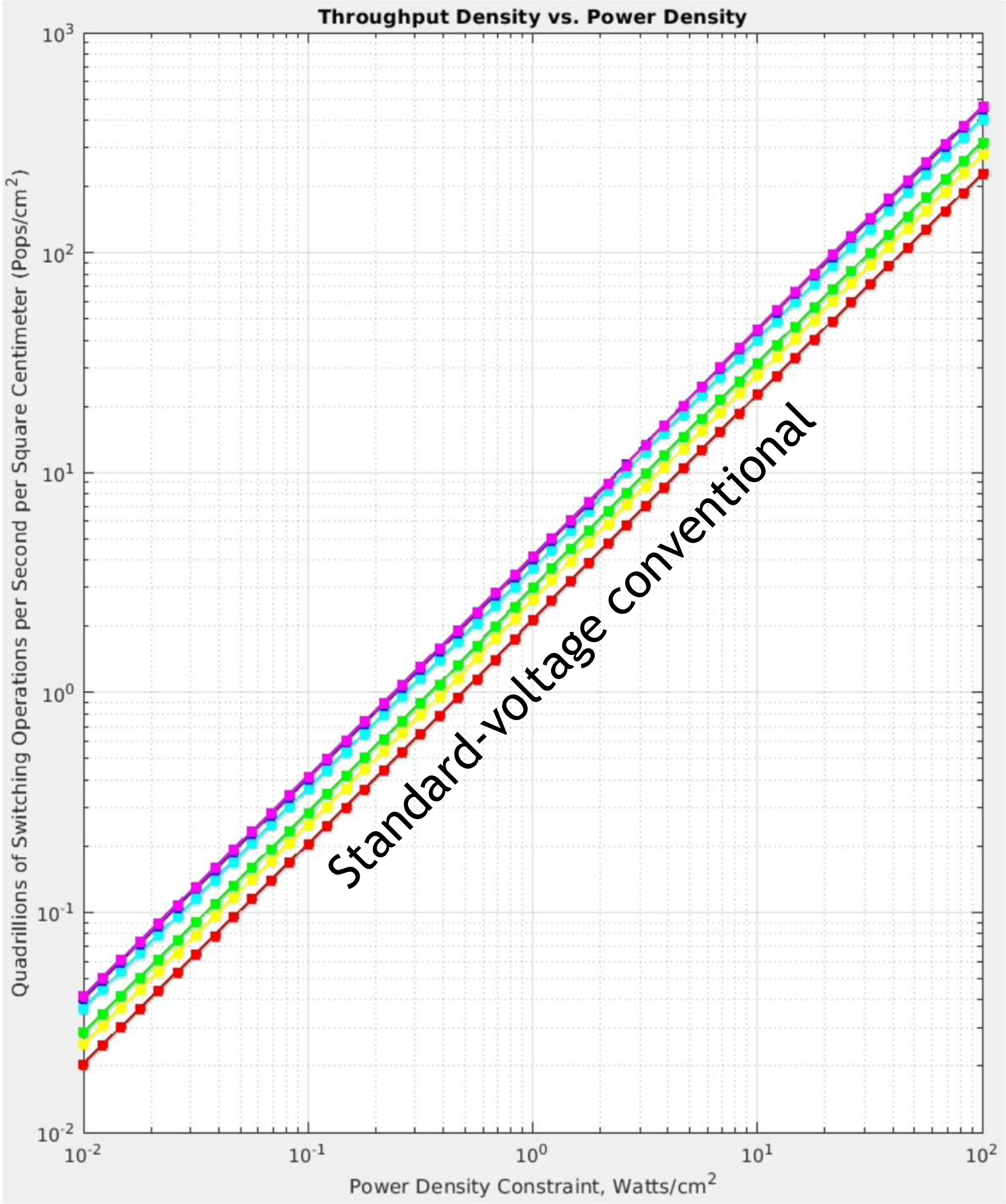
Here, we maintain leakage power at no more than 10% of total power by decreasing average gate density as needed.

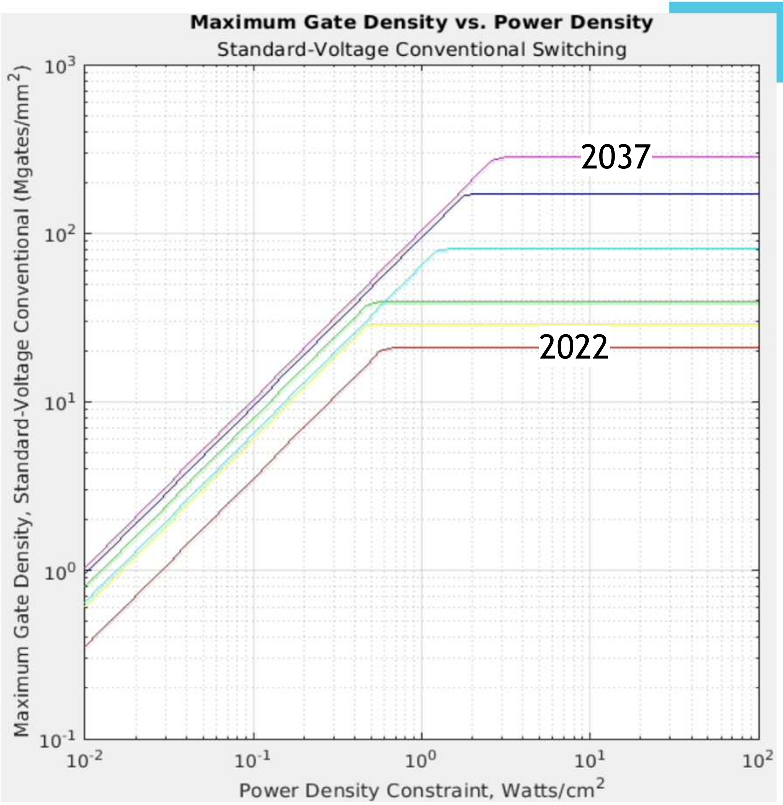
With conventional switching at standard voltages, throughput falls ∝ power density, as expected – since energy dissipation per switching event is a constant.

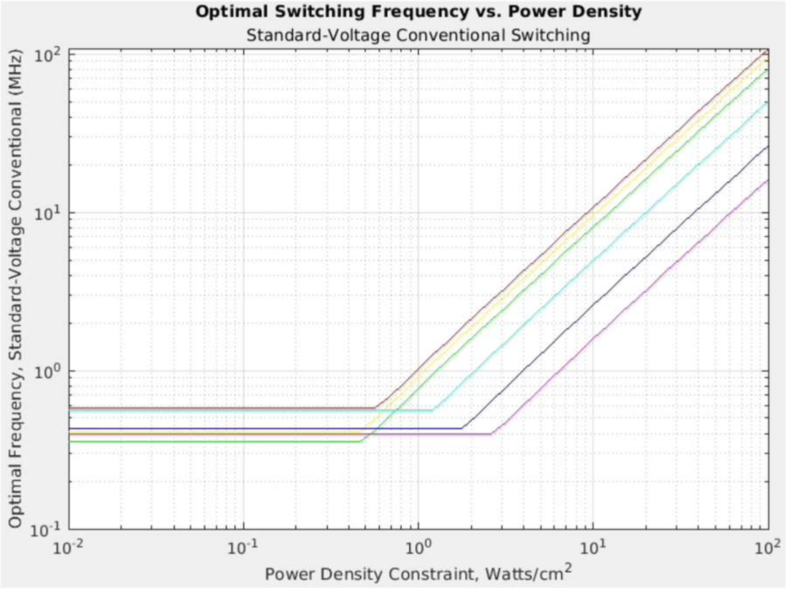
- Note that at max density of active gates, switching frequency can be no more than ~1 MHz/W in '22!

Note also that throughput improves by only about 2× between 2022 and 2037!

- Because, see slide 14.







Voltage-Optimized Conventional Switching

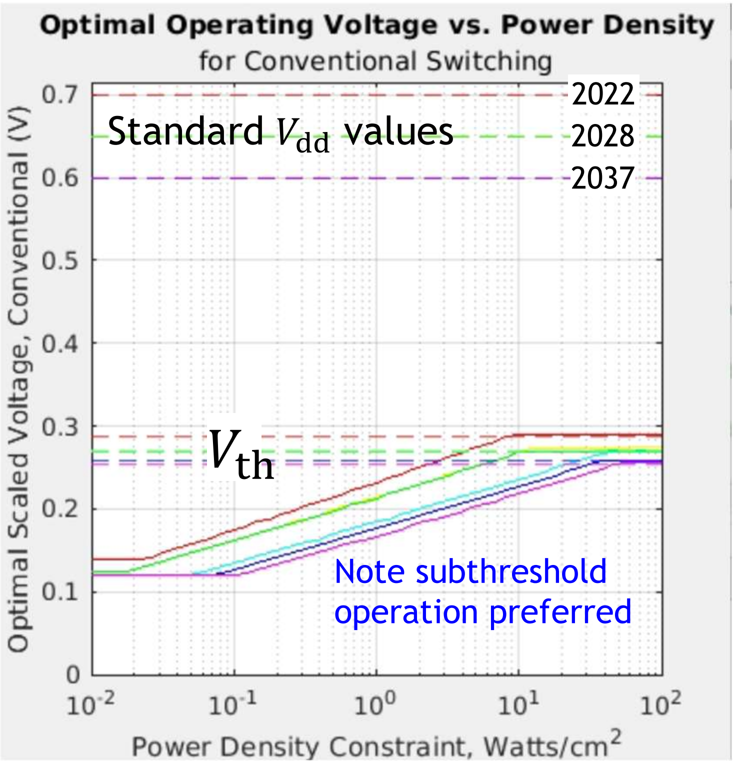
Note optimal voltages for maximum throughput density start near threshold, and trend subthreshold at lower power levels.

- End up at roughly ½ of threshold level.

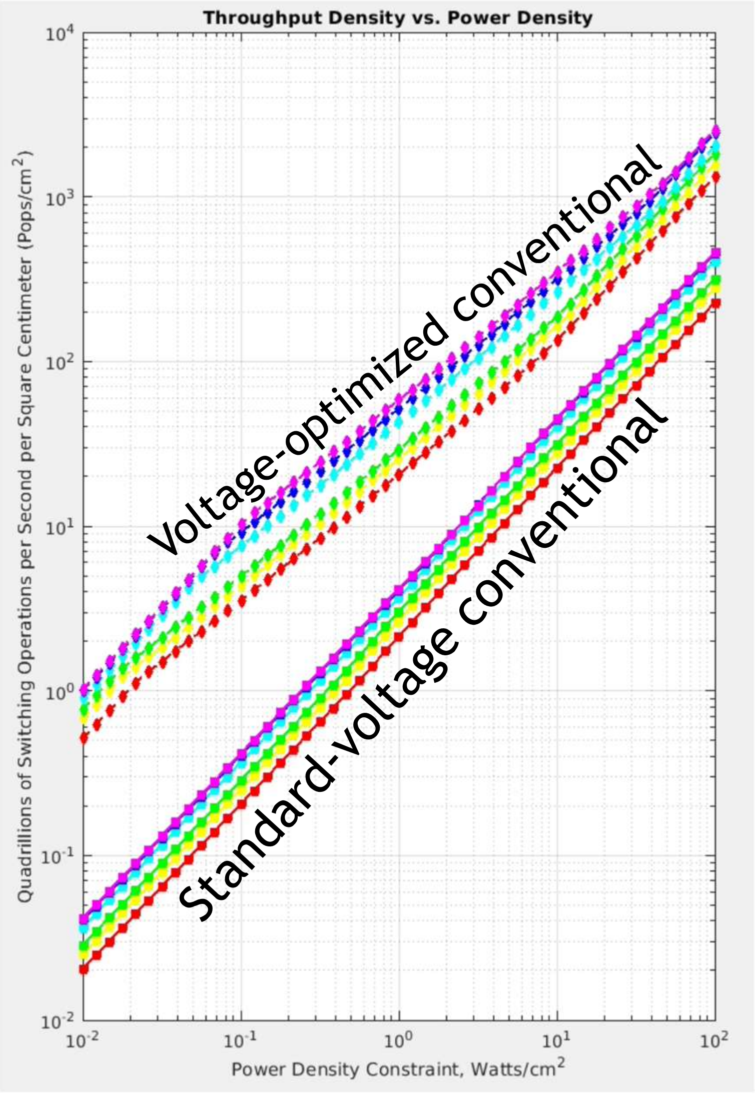
Because of low V_{dd}, leakage power is greatly reduced, and doesn't start to limit max gate density until very low power density levels.

Maximum frequency at max gate density also improves vs. higher-V_{dd}, and moreso as the power limit & switching voltage decreases.

- ~24.6× throughput boost at low power per die

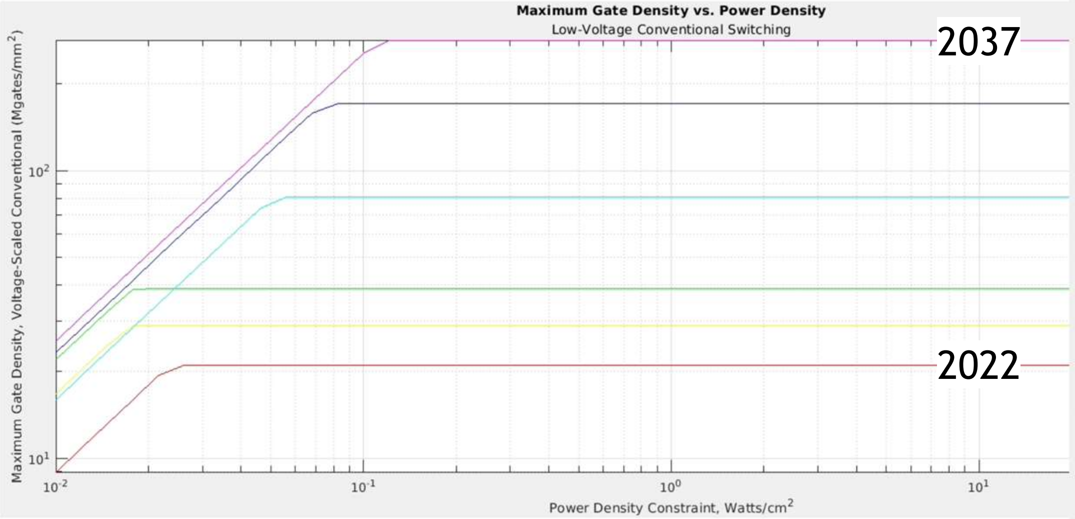


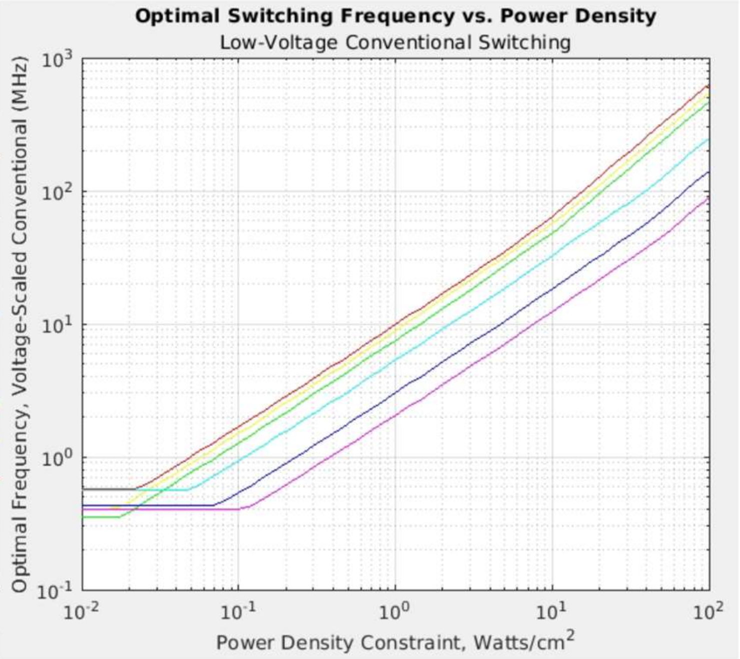
Note subthreshold operation preferred



Voltage-optimized conventional

Standard-voltage conventional





Standard-Voltage Adiabatic Switching

Here, we make sure that leakage is no more than 50% of total power

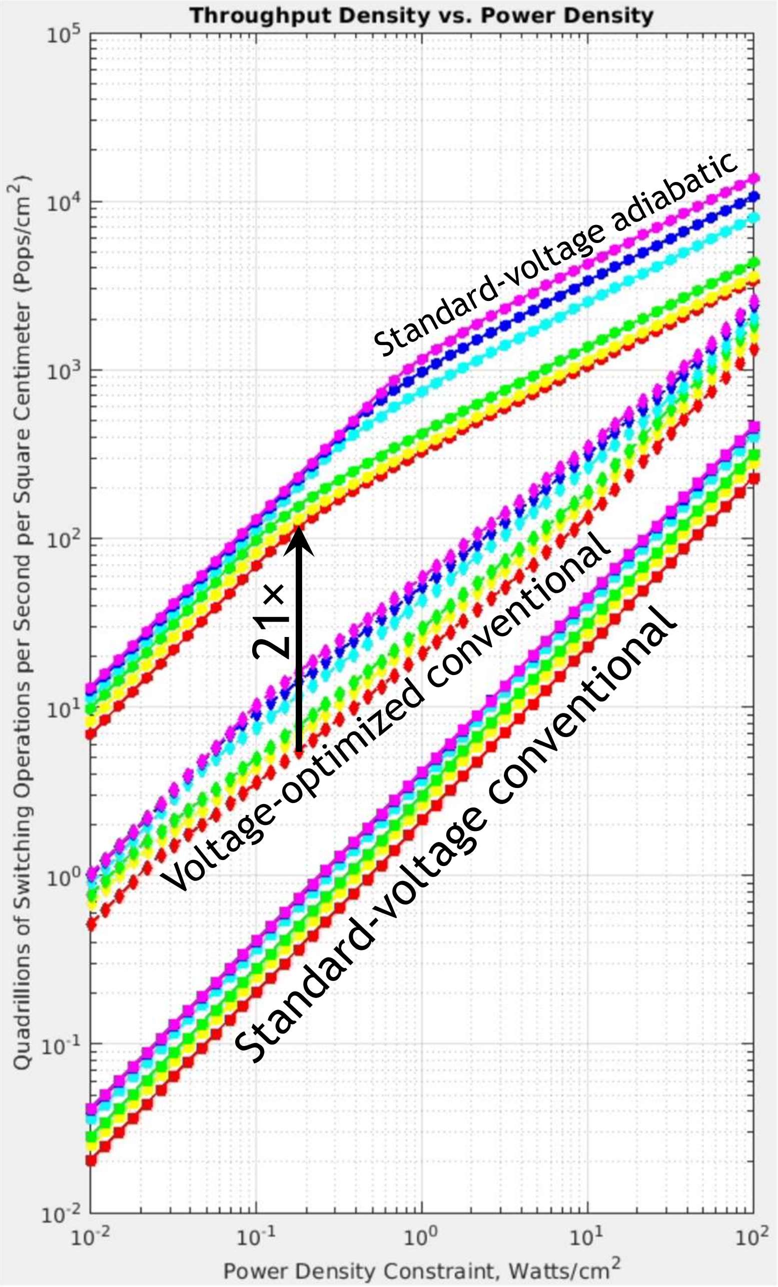
- Because that is the point of theoretical maximum efficiency for adiabatic switching

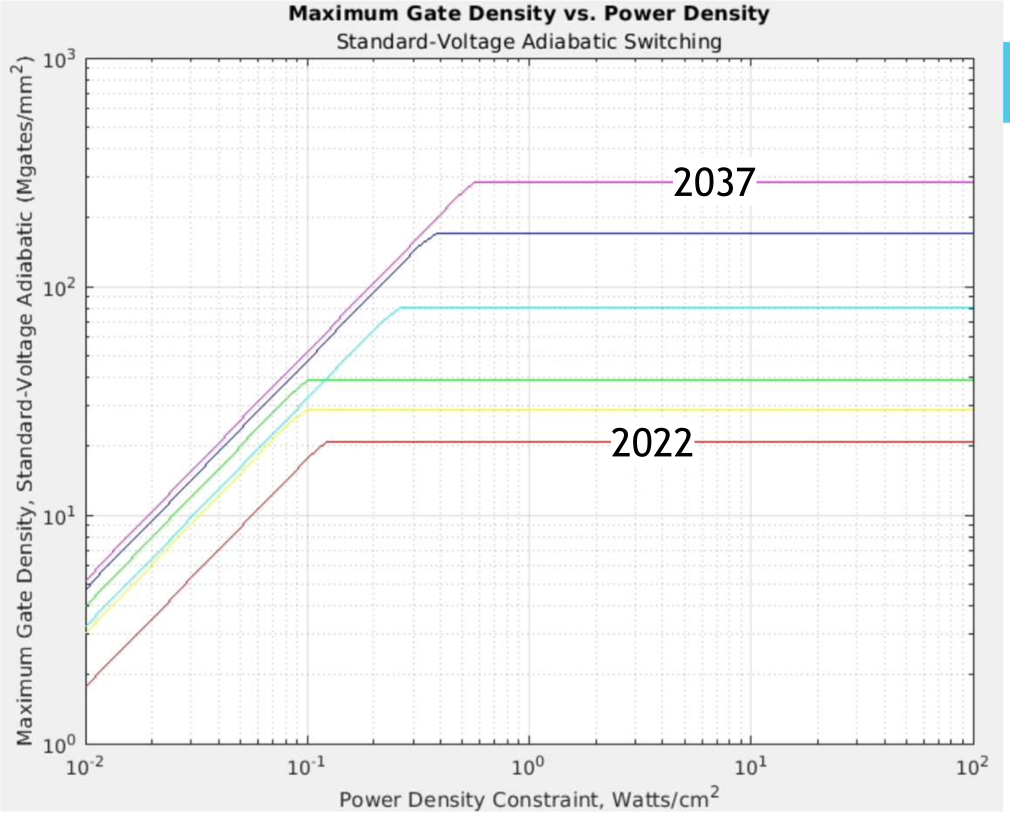
Note adiabatic frequency vs. power curves are ½ as steep as standard-voltage conventional.

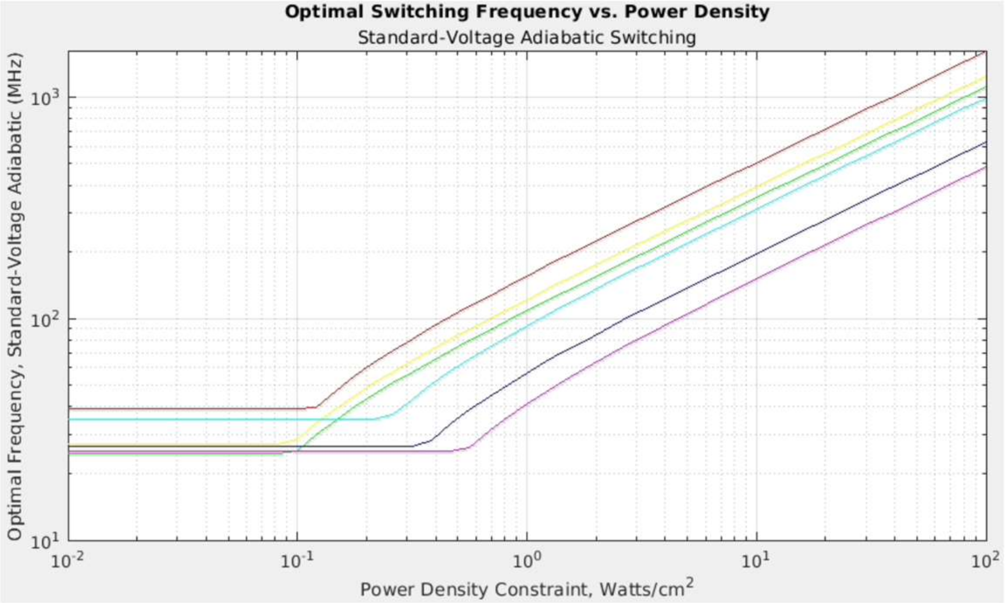
- Increasing adiabatic advantage at low per-die power densities!

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/cm^2 in 2022)







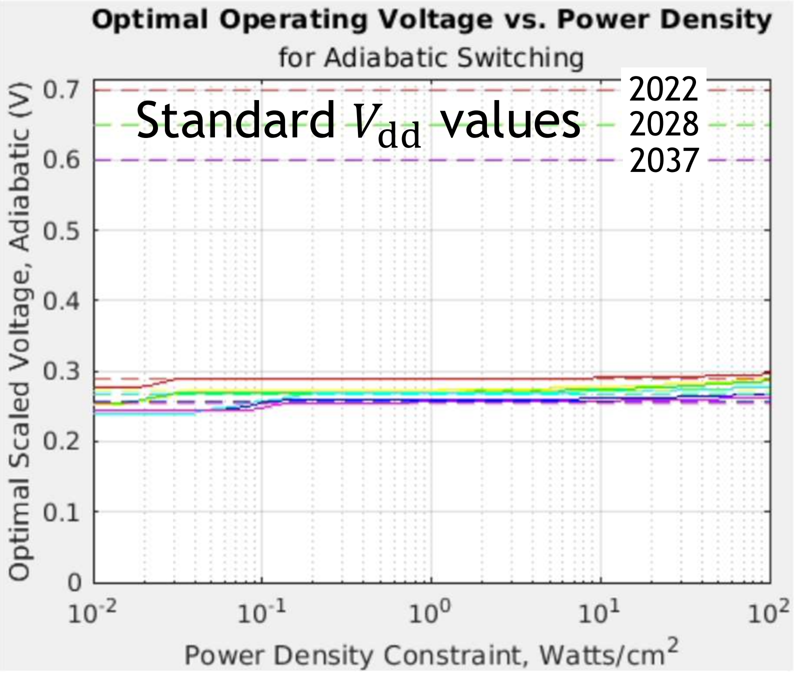
Voltage-Optimized Adiabatic Switching

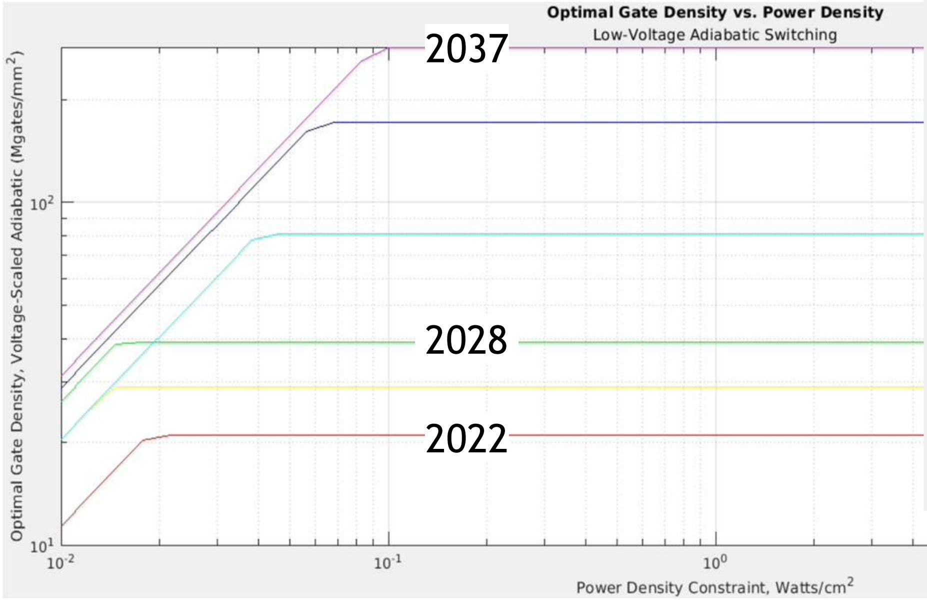
This time, the optimal voltages end up near-threshold but not significantly subthreshold

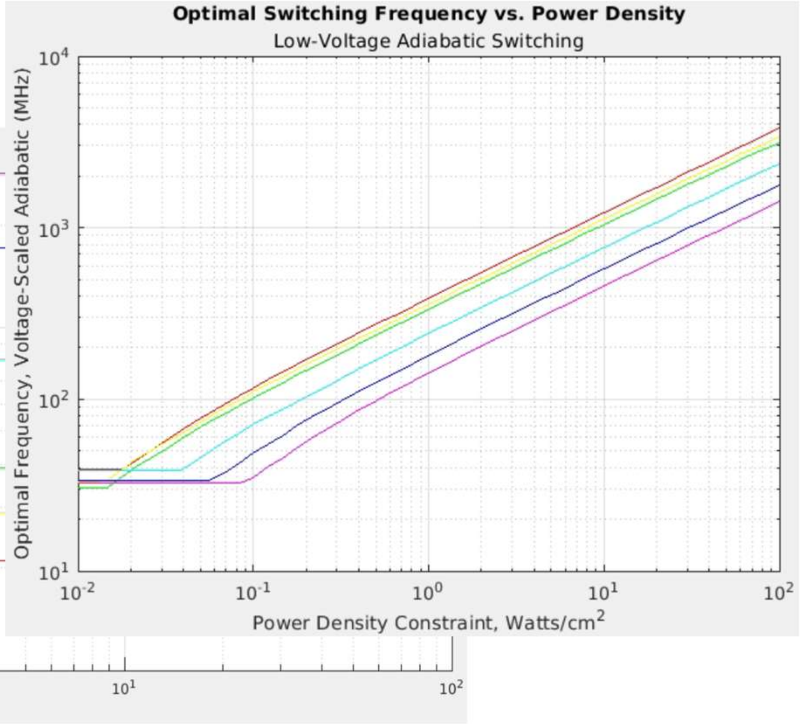
- Note this improves noise immunity vs. optimized conventional CMOS

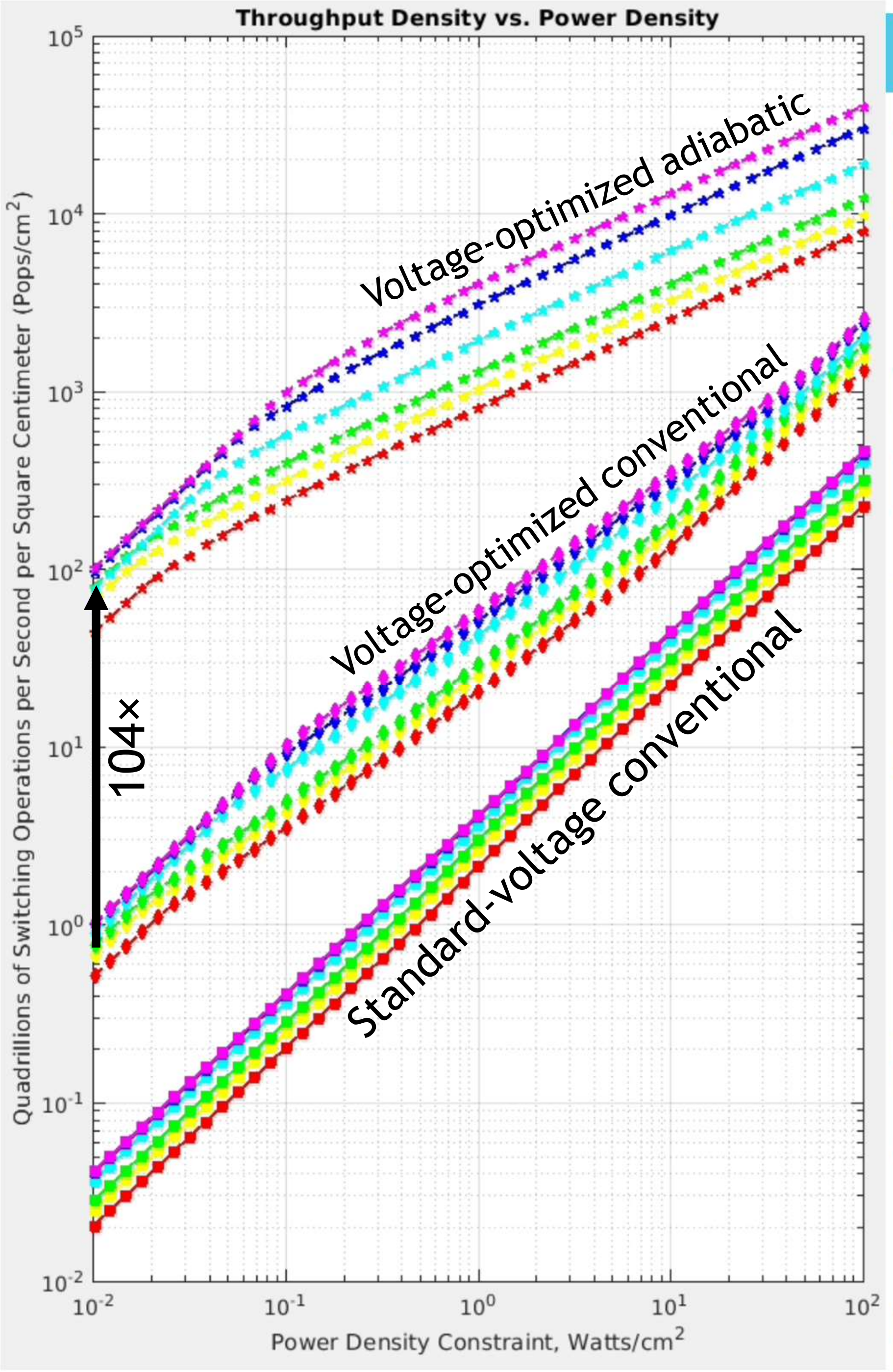
Adiabatic scaling advantage extends farther before limited by leakage.

- Maximum boost vs. conventional CMOS is now 104× (in 2028) at low power-per-die levels









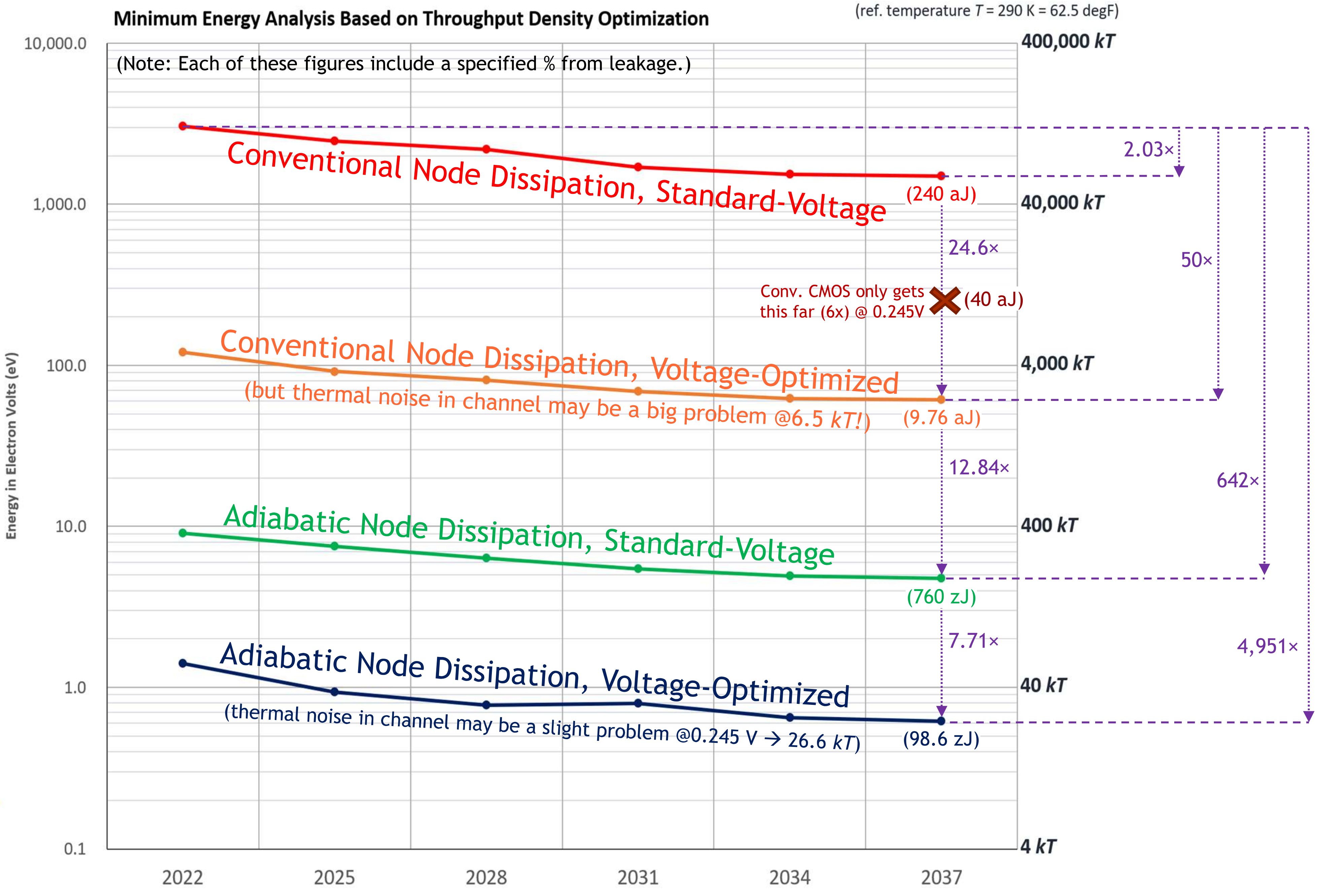
Summarizing the Preliminary Energy Efficiency Results from our Throughput-Density Maximization Study

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!

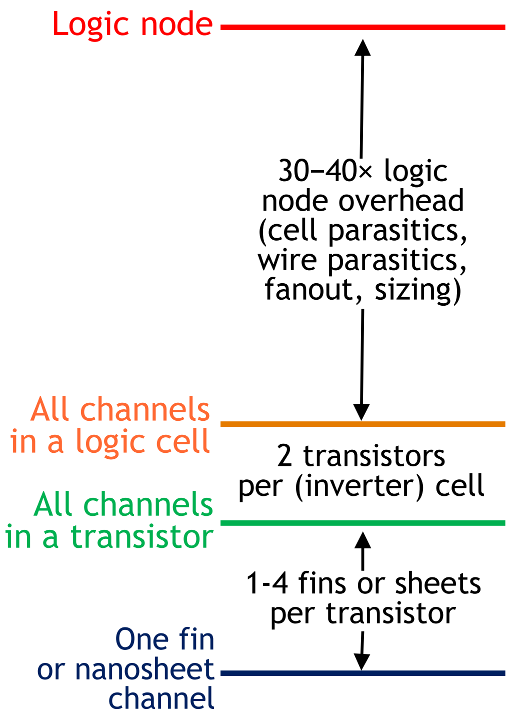
Adiabatic beats conventional by ~405× at opt. adia. voltage (0.245 V) if it's achievable.



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!



Logic node

30–40× logic node overhead (cell parasitics, wire parasitics, fanout, sizing)

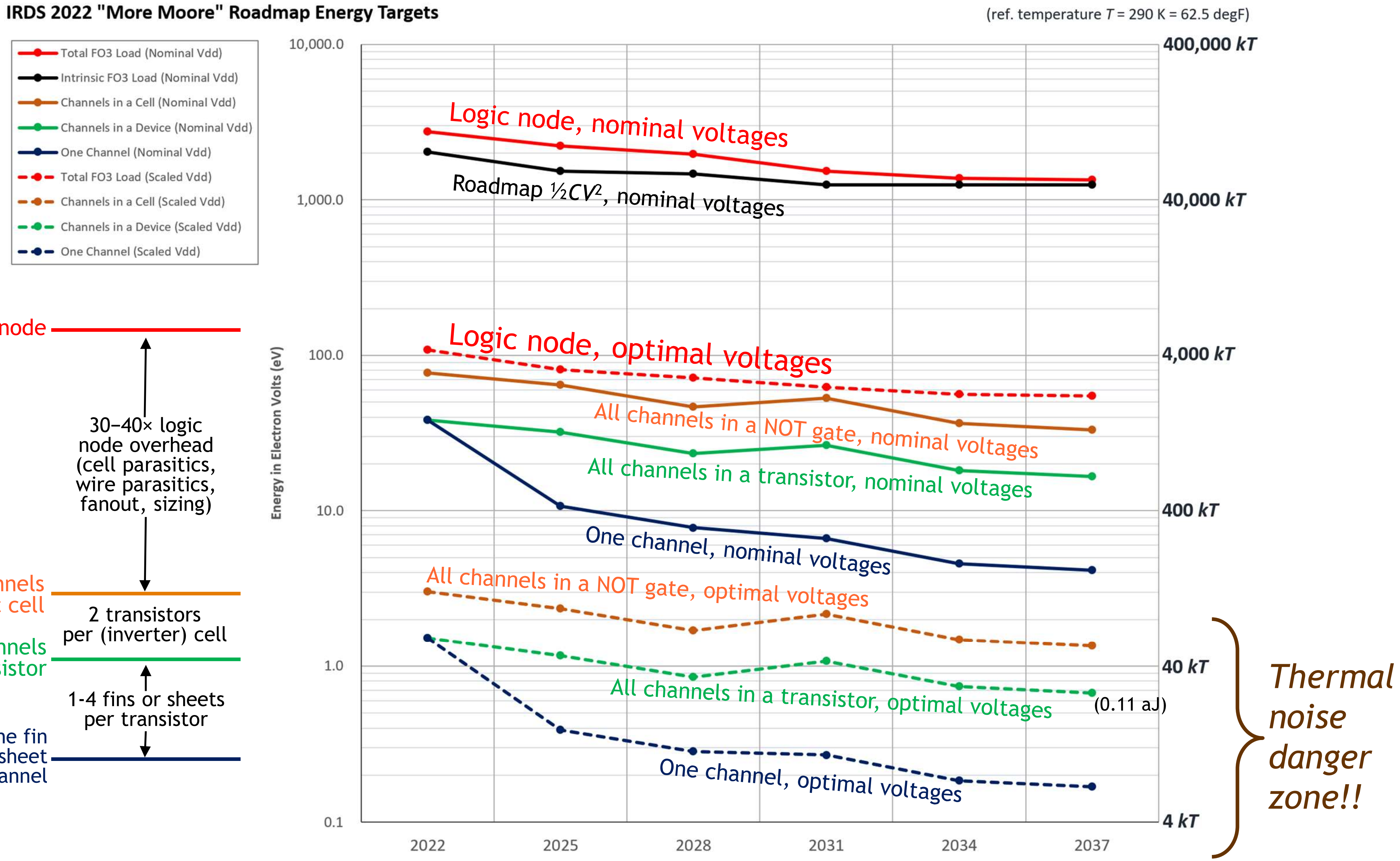
All channels in a logic cell

2 transistors per (inverter) cell

All channels in a transistor

1-4 fins or sheets per transistor

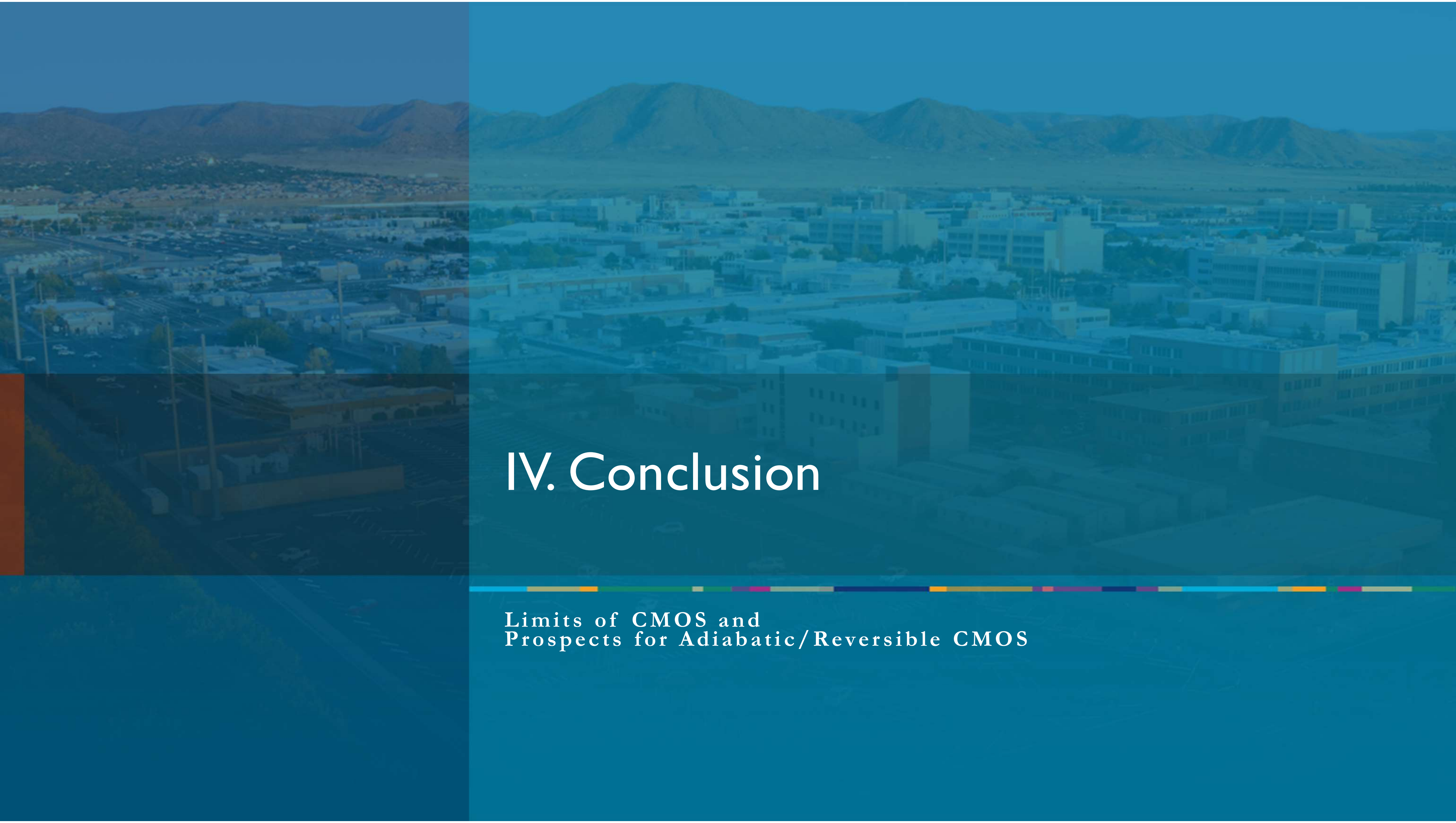
One fin or nanosheet channel



IRDS 2022 "More Moore" Roadmap Energy Targets

(ref. temperature T = 290 K = 62.5 degF)

Thermal noise danger zone!!



IV. Conclusion

Limits of CMOS and
Prospects for Adiabatic/Reversible CMOS

Conclusion & Next Steps

Preliminary conclusions from the present study to date:

- Conventional CMOS is fast approaching fundamental limits from thermal noise!
  - Only **~2–12×** 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 **~100×** vs. end of conventional CMOS (even including subthreshold CMOS!), or **~400×** 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