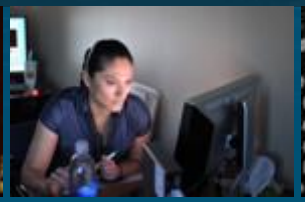
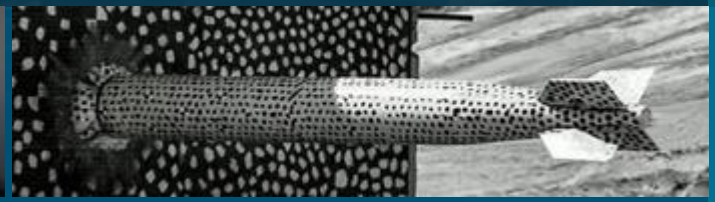
IEEE International Conference on Rebooting Computing (ICRC 2022)
San Francisco, California, USA, December 8-9, 2022
Panel on: "Continuing the Progress in Energy Efficient Computing"

Sandia National Laboratories

Continued Efficiency Scaling of General Digital Compute via Reversible Computing





Thursday, December 8th, 2022

Michael P. Frank, Center for Computing Research

with Rupert Lewis (Quantum Phenomena Dept.)

Approved for public release, SAND2022-17020 C

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

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.

Energy Efficiency of Computing: The Past

Total intensity of computing worldwide has increased by ~6 orders of magnitude over the last 30 years!

- Order ~1,000× or so every 15 years (~10× every 5 years, ~100× every 10 years).

Example calculation: As a rough proxy for global aggregate computing intensity, let's consider all the machines on the TOP500 list taken together:

- First (June 1993) edition:
  - Aggregate max sustained performance = 1.123 Tflop/s
- Most recent (Nov. 2022) edition:
  - Aggregate max sustained performance = 4,864,000 Tflop/s (4.8 Eflop/s)
- Overall increase of 4.33M× over 29.5 years
  - Average increase of 68.1% per year = 2,400× per 15 years

Arguably, the increase in compute has enabled much of the economic growth in the past 30 years.

Would this increase have been possible *without* comparable increases in energy efficiency? → **NO.**

- Energy efficiency (not increased energy consumption) accounts for the vast majority of increase.

Can we hope for comparable increases in total compute (& associated economic growth) without similar increases in energy efficiency over the *next* 30 years (to ~2050)? → **ABSOLUTELY NO.**

- Computing *already* accounts for a significant fraction of total electric power usage!

Energy Efficiency of Computing: The Present

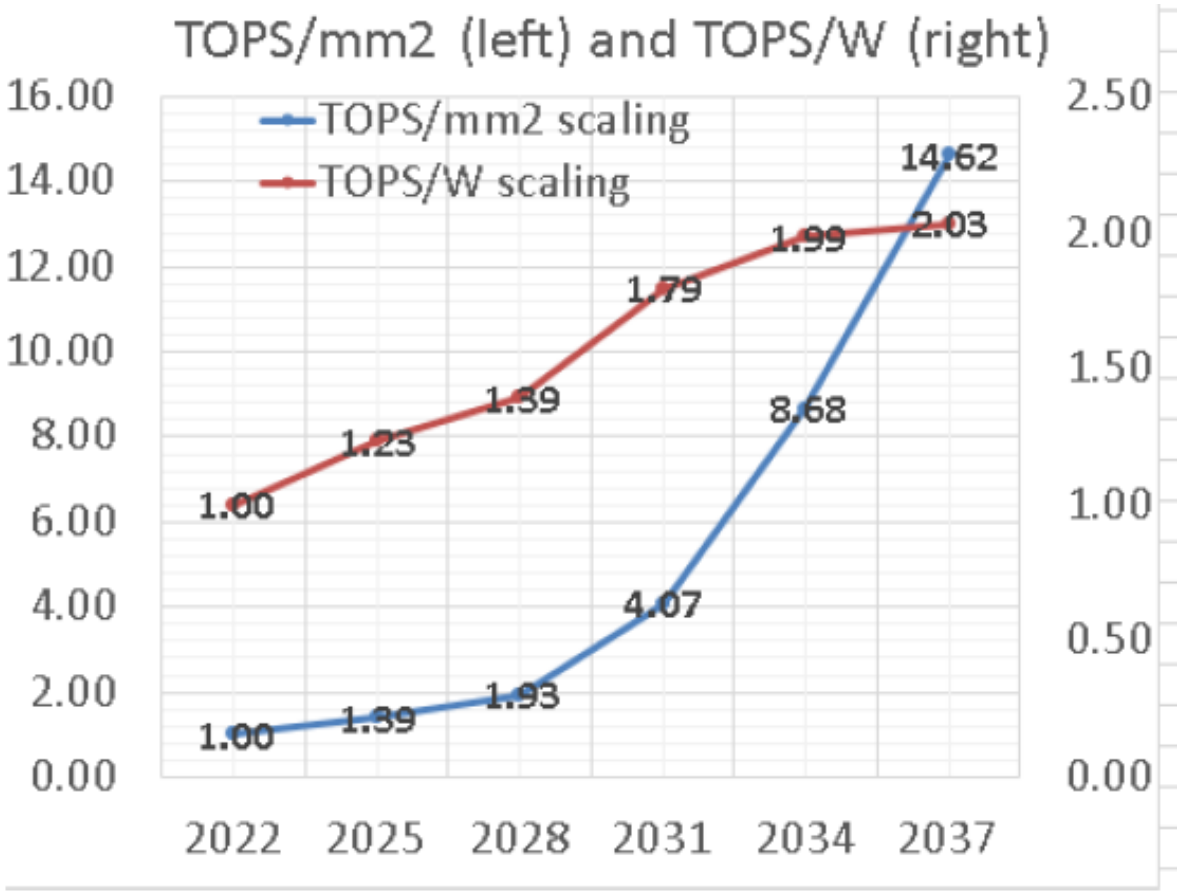
What is the outlook given "business as usual" CMOS technology development looking forward?

- Unfortunately, it's not very good!!
  - Roadmap has low-level energy efficiency increasing by only ~2× over next 15y!

By 2037, we are very close to fundamental limits:

- Δ channel charge per nanosheet is only **~14 electrons** (or **~3 e^-** for low-voltage operation).
- Δ ½CV^2 gate energy per quantum channel is **~80 *kT***. At low voltages, total gate energy for an entire standard logic cell is only **~50 kT**.

YEAR OF PRODUCTION	2022	2025	2028	2031	2034	2037
LOGIC DEVICE GROUND RULES						
M0 pitch (nm)	24	20	16	16	16	16
Gate pitch (nm)	48	45	42	40	38	38
Lg: Gate Length - HP (nm)	16	14	12	12	12	12
Lg: Gate Length - HD (nm)	18	14	12	12	12	12
FinFET Fin width (nm)	5.0					
Footprint drive efficiency - finFET	4.21					
Lateral GAA vertical pitch (nm)		18.0	16.0	16.0	15.0	14.0
Lateral GAA (nanosheet) thickness (nm)		6.0	6.0	6.0	5.0	4.0
Number of vertically stacked nanosheets on one device		3	3	4	4	4
LGAA width (nm) - HP		30	30	20	15	15
LGAA width (nm) - HD		15	10	10	6	6
LOGIC DEVICE ELECTRICAL SPECS						
Power Supply Voltage - Vdd (V)	0.70	0.65	0.65	0.60	0.60	0.60
Vt,sat at Ioff=10nA/um - HP (mV)	156	165	165	164	156	154
Vt,sat (mV) at Ioff=100pA/um - HD (mV) [3][4]	288	271	268	268	258	255
Cch,total (fF/um2) - HP/HD [8]	34.52	34.52	38.35	38.35	38.35	38.35
Energy per switching [CV2] (fj/switch) - FO3 load, HP	0.65	0.49	0.47	0.40	0.40	0.40
LOGIC CELL ELECTRICAL FoMs						
Total loading capacitance (ff) - HP	5.90	5.65	5.35	4.76	4.00	3.86
Total loading capacitance (ff) - HD	1.79	1.69	1.49	1.36	1.23	1.20
Dynamic power at 1GHz clock (mW) - HP	2.89	2.39	2.26	1.71	1.44	1.39
Dynamic power at 1GHz clock (mW) - HD	0.88	0.71	0.63	0.49	0.44	0.43



(Data & chart from IRDS 22 More Moore chapter)

Energy Efficiency of Computing: The Future

Because low-level energy efficiency of CMOS is only increasing ~2× over the next 15 years (by 2037), power-density-limited throughput density will also only increase by ~2× over this period!

- You can scale a little better with low-voltage operation, but not much…

But, fully **adiabatic switching** (w. quasi-trapezoidal AC supplies) can (in principle) improve the raw switching throughput density (per die) by up to **~100×** compared to conventional switching (even low-V).

- Realizing these kind of benefits requires (at least local) use of **reversible computing** design principles.

But, reversible design imposes its own overheads, and there are *many engineering challenges* that will need be addressed in order to produce practical adiabatic/reversible designs with substantial benefits.

- Design of high-Q resonators and power-clock distribution networks.
- Load balancing: Minimizing random variations in cycle-to-cycle energy.
- Achieving increased *dynamic* (AC supply RMS) power density.
- Minimizing overheads of reversible (hardware) algorithms for target functions.
- EDA tool development to support adiabatic/reversible designs.

If we want to continue improving the energy efficiency (throughput per Watt) of digital compute at comparable to historical rates, we ***have to*** start seriously tackling the challenges of adiabatic/reversible design.