Large and small: The problems of scales in semiconductor electronics

Sandip Tiwari, stiwari@iitk.ac.in, st222@cornell.edu

In the use of nanoscale(10^{-9} *nm*)-sized structures integrated now to terascale (10^{12}) at the chip level and exascale (10^{18}) in supercomputing and in the cloud, the design, the making, and the implementation in electronics is *statistical information mechanics*.

Deterministic as in general purpose and non-deterministic computation as in AI/ML are under *different constraints*.

The subject range and the importance of the problem lends itself to very interesting and productive paths, but also lot of fallacies.

1. Large and small: The problems of scales in semiconductor electronics
2. Non-Turing machines: Stochastic and probabilistic learning circuits
3. Physics-guided AI/ML: Why, how and usage

4. Cultures: Science, engineering, interdisciplinarity and the fallacy of Ockham's razor
5. Semiconductors: Lessons from the past and what it says for semiconductor manufacturing

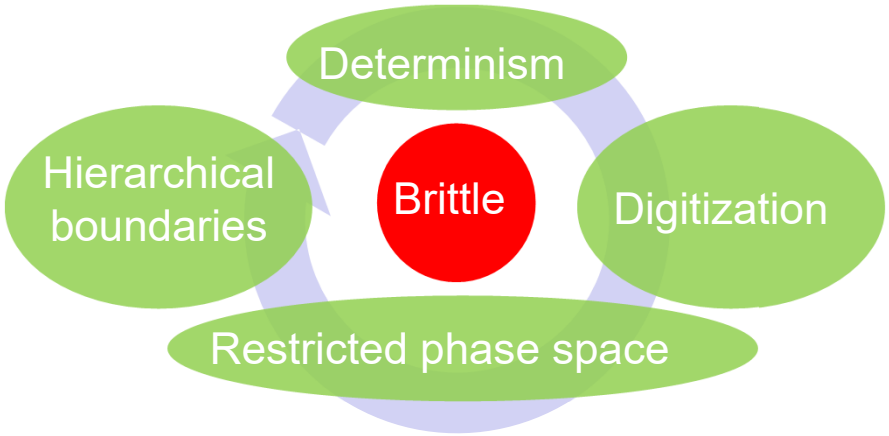
This talk is the first of 5 e*xploring the information engine and technology in our world.*

The first 3 are technical.
The last 2 build on the technical but are broader exploration using my mind's eye
by *stepping back to view the lessons of past, science as a social force, the nature of its pursuits and use, and its lessons for the future in a multipolar world of many different political system and different people needs.*

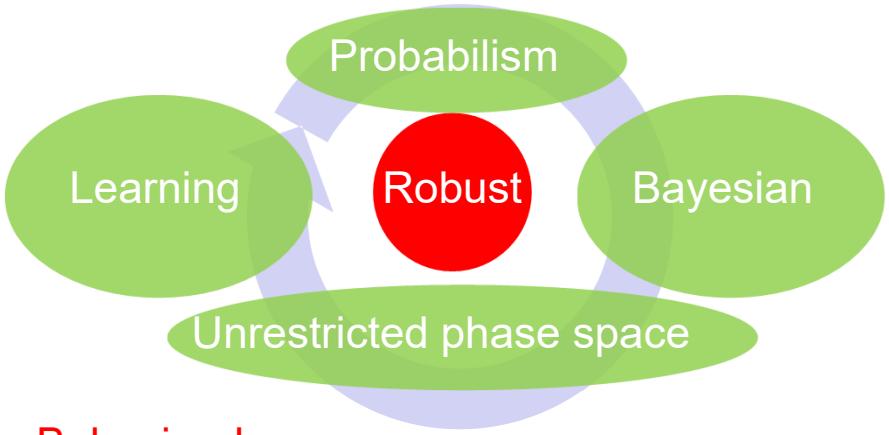
https://www.iitk.ac.in/scdt/Sandip_Tiwari_Kanpur_Lectures.html

1971	1980	2023
Silicon bipolar transistor	digital → HF analog	SiGe bipolar transistor
ECL circuits	nMOS circuits	CMOS circuits static and dynamic
Some nMOS circuits (mainly memory)	CPUs	CPUs GPUs Tensor engines
first cpu Intel 4004 4 *b*, 2.25 x 10^2 *trx*	Motorola 68000 16/32 *b*, 6.8x10^4 *trx*	Apple A16 (4 *nm*), 1.6 x 10^{10} *trx* Navi 31 (5 *nm*), 5.8 x 10^{10} *trx* Biren BR100, 7.7 x 10^{10} *trx* Hopper H100, *3 TB/s*, 7.7 x 10^{10} *trx* *60 TFLOPs FP64 tensor; 700W*
Determinism		Determinism & Indeterminism using deterministic resources → ? Inherent indeterminism

Two universes



Exact in Boolean sense

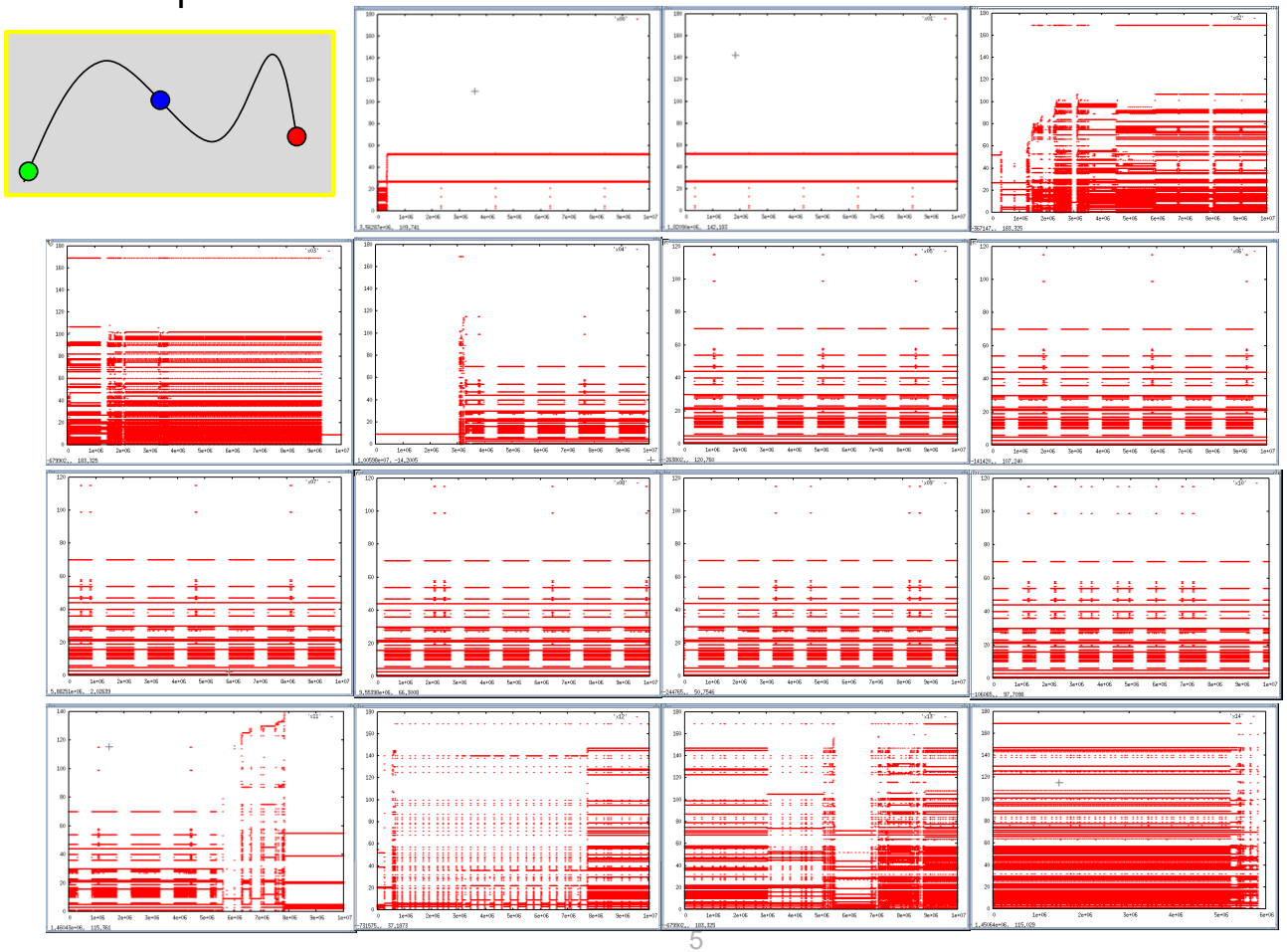


Behavioral

Information is also connectivity.
Involves learning,
inexactness and incompleteness.

Phase Space of data machines

Deterministic computing machines sample only a specific small subset of states responding to data, instruction set and precision.



Data ≠ Information

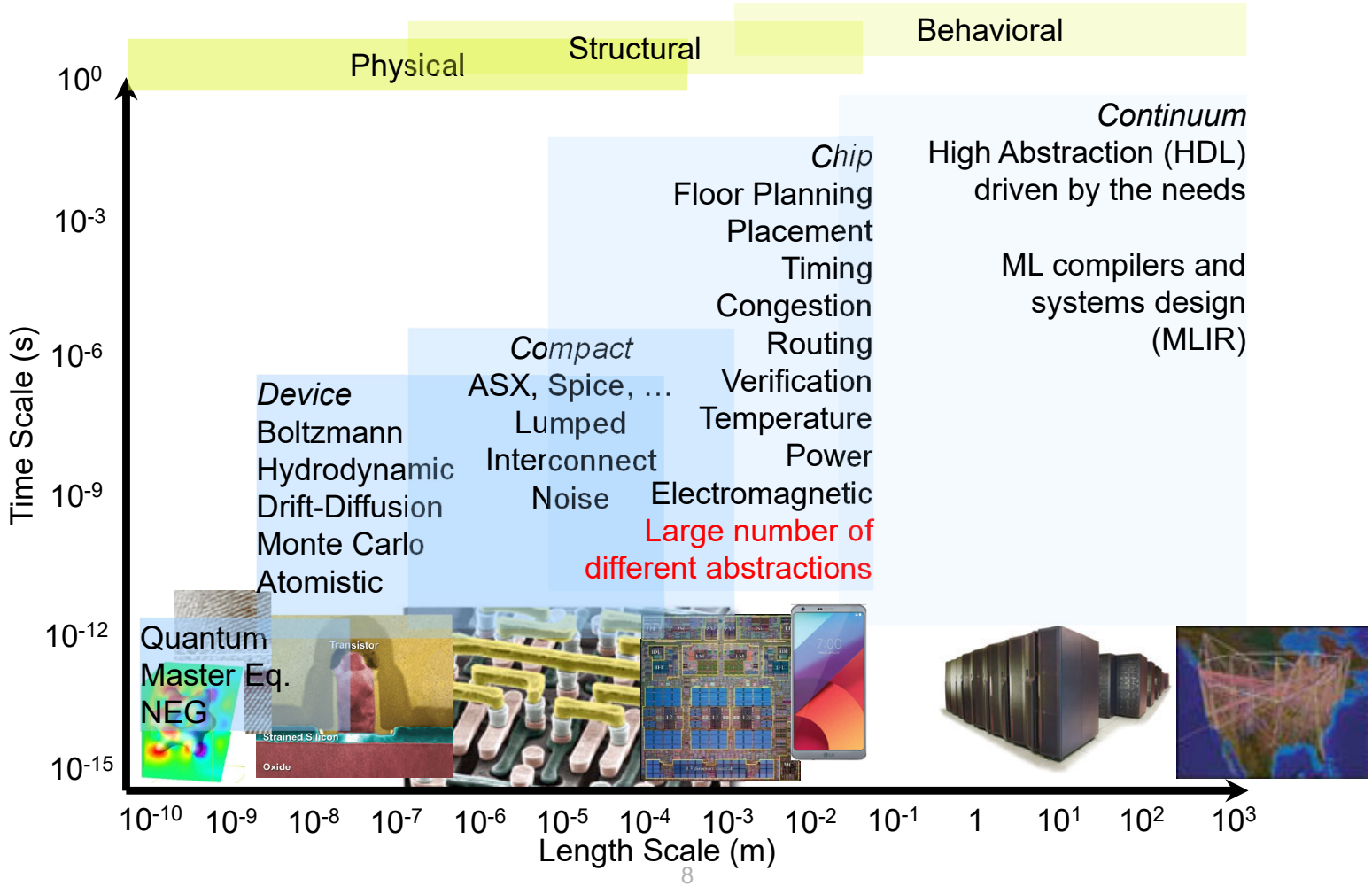
Olshausen (2008)

Data ≠ Information

Olshausen (2008)

1 piece of data (black-fill) reveals the information.
Context was important. Algorithmic.
Information and data are not equivalent

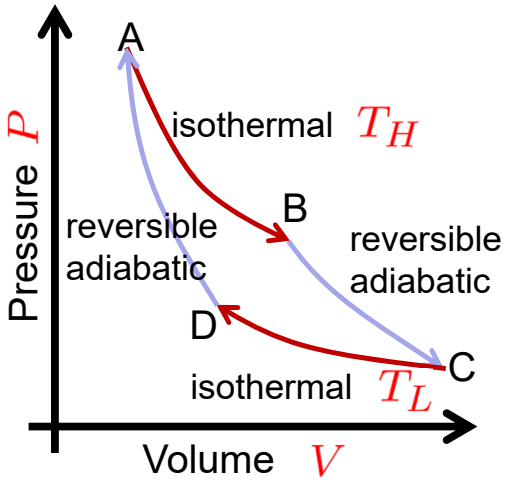
Scale Changes

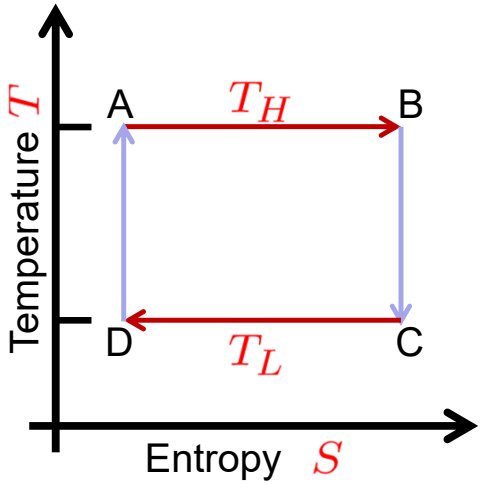


Energy efficiency

Nicolas Léonard
Sadi Carnot

Thermomechanical:





Free energy $đU = đW + đQ$ Heat

Useful work

Work is <u>useful</u> work - for mechanical, somewhat definable, though it too leads to inconsistencies because of path dependence.

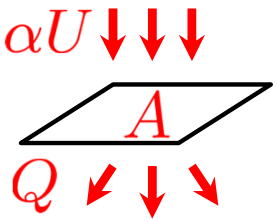
For information, not quite even that clear.

In Shannon view, it means the manipulation of the “work” content embedded within the data.

But, this is also not information or knowledge, which is the network of connections within the data.

Emergence: electronics power and speed

Steady state:



$$\tau = \frac{\alpha U}{QA}$$

- α: activity factor
- U: energy per operation
- A: cross-section area of heat removal
- τ: time constant of energy consuming operation
- Q: density of heat removal

1D Heat Spreading:
Array regions

$Q = 10^2\ W/cm^2$

$\Rightarrow \tau = 5\ ns$

3D Heat Spreading:
Some logic regions

$Q = 10^5\ W/cm^2$

$\Rightarrow \tau = 5\ ps$

Determinism, Rates, Errors & Power

Trapped in a macrostate:

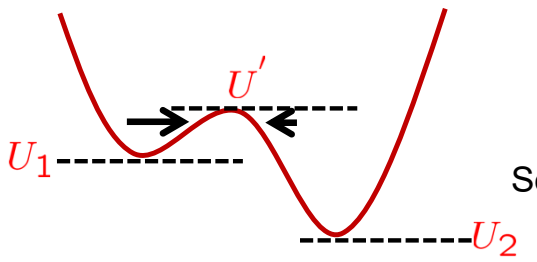
$$T_s \approx \tau_s \exp(-\frac{\Delta U}{kT})$$

$$\tau_s = 10^{-12} \ s, \ \ T_s = 10 \ \ yr \Rightarrow \Delta U \approx 50 k_B T$$

Thermal will dominate

τ_s has thermal effects too

Transitions from a macrostate:



ratio of forward to backward: $\exp\left(-\frac{U_1 - U_2}{k_B T}\right)$

So, energy expended per step: $k_B T \ln r = U_1 - U_2$

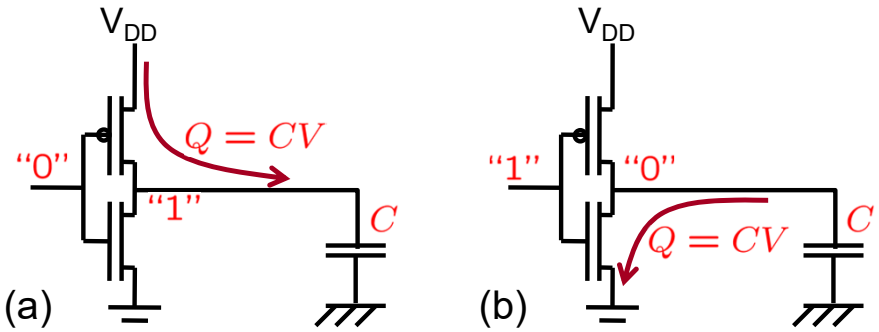
Error rates of a macrostate:

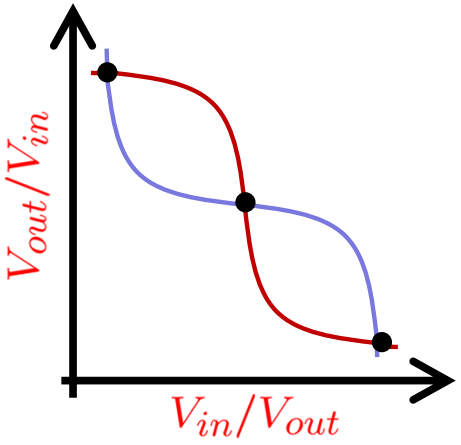
Failure probability: $\mathfrak{q}(= 1 - \mathfrak{p}) \propto \exp\left(-\frac{\Delta U}{k_B T}\right)$

Errors are exponentially related to energies.

Real world processes controlled by thermodynamics.

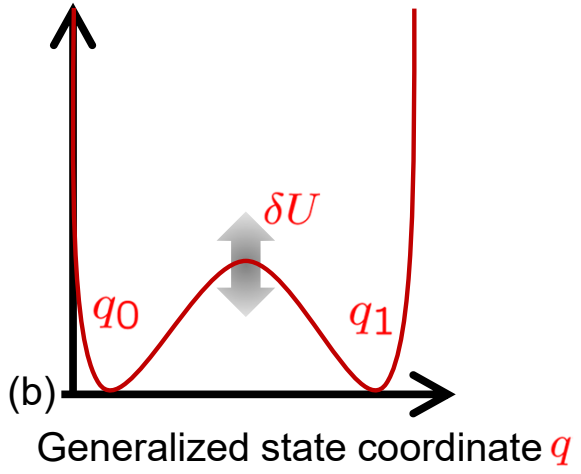
Static CMOS





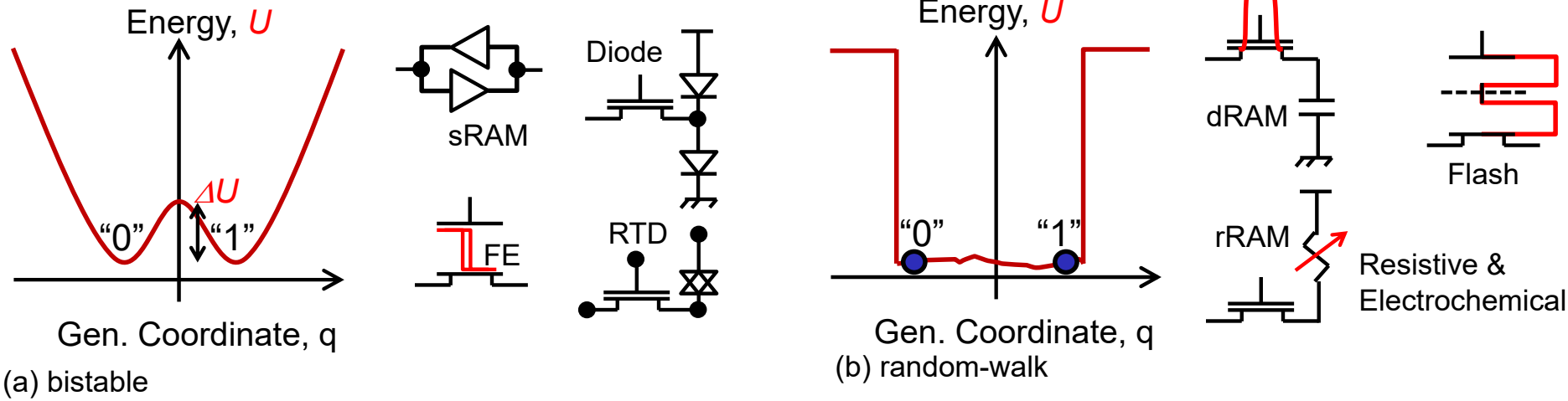
In the process of state change, static CMOS just dumps all the energy to the ground reservoir.

Not-recoverable, and only in few instances does one recover it since speed (non-reversibility and non-adiabaticity) is desired.



Memories

Bistable

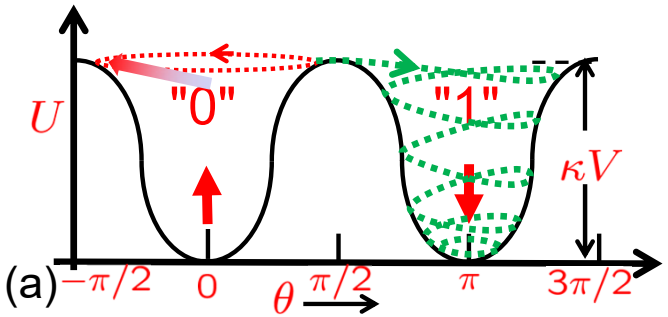


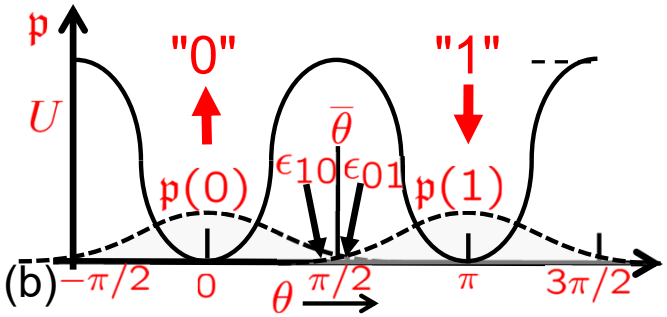
(a) bistable

Random walk

(b) random-walk

Spin-torque and magnetic memories





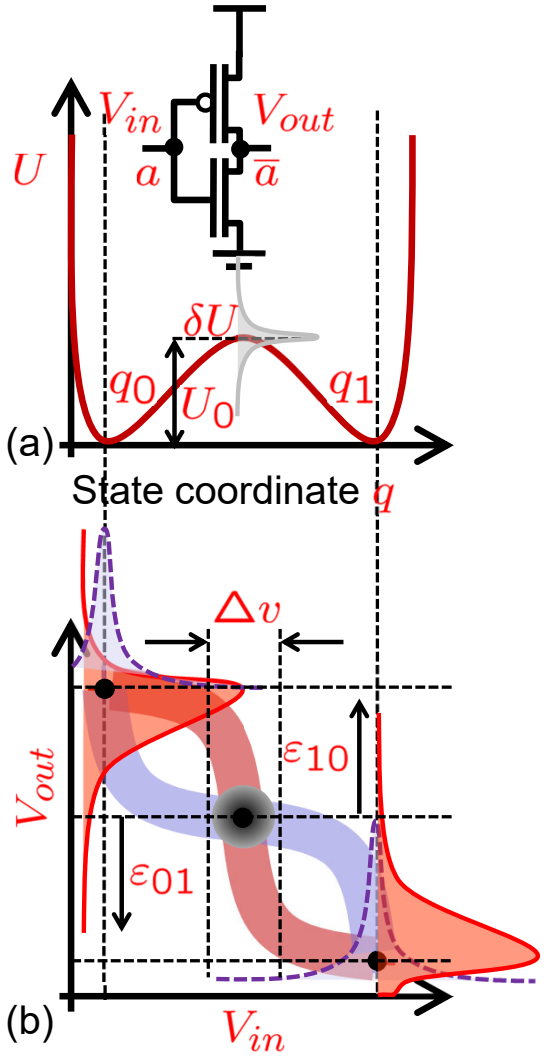
$$\frac{1}{\tau_{N,B}} = \mathcal{A}_{N,B} \exp\left(-\frac{\kappa\Omega}{k_B T}\right) \qquad \mathcal{A}_{N,B} = (\gamma H_{co}/2)(\kappa\Omega/k_B T\pi)^{1/2}$$

Terabyte with 1 in 10 years:

$$\tau_{N,B} \approx 2.5 \times 10^{21}\ s$$

$$\Delta U = \kappa\Omega \approx 75 k_B T$$

$$\mathfrak{q}(=1-\mathfrak{p}) \propto \exp\left(-\frac{\Delta U}{k_B T}\right)$$



Static CMOS

$$\mathfrak{p}(0) = \frac{1}{\sqrt{2\pi}}\left[1 - gS(v, v_x, \Delta v)\right] \exp\left[-\frac{(v - v_{inL})^2}{2k_BT}\right]$$

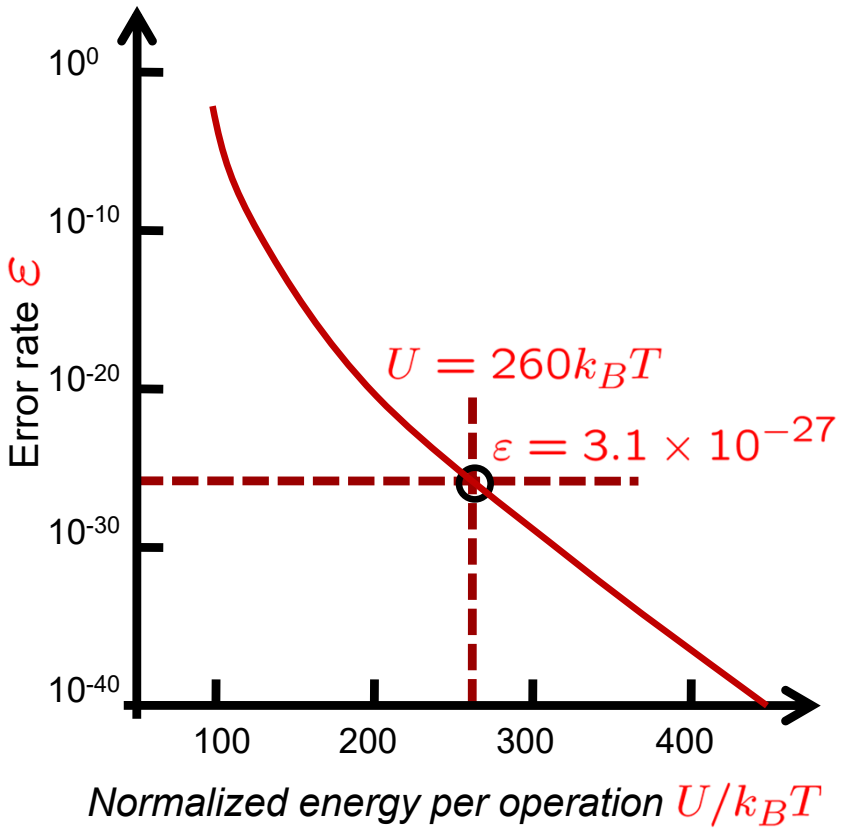
$$S(v, v_x, \Delta v) = \frac{1}{1 + \exp\left[g(v - v_x)/\Delta v\right]}$$

$$\mathfrak{p}(1) = \frac{1}{\sqrt{2\pi}}\left[1 - gS'(v, v_x, \Delta v)\right] \exp\left[-\frac{(v - v_{inH})^2}{2k_BT}\right]$$

$$\begin{aligned}\varepsilon &= \frac{1}{2}(\varepsilon_{01} + \varepsilon_{10}) \\ &\approx \frac{2}{\sqrt{2\pi}} \int_{v_x}^{\infty} \left[1 - gS(v, v_x, \Delta v)\right] \\ &\quad \times \exp\left[-\frac{(v - v_{inL})^2}{2k_BT}\right] d\left(\frac{v}{2\sqrt{k_BT/C}}\right).\end{aligned}$$

Model error calculation (Gaussian, thermal)

1 device failure in 10 years



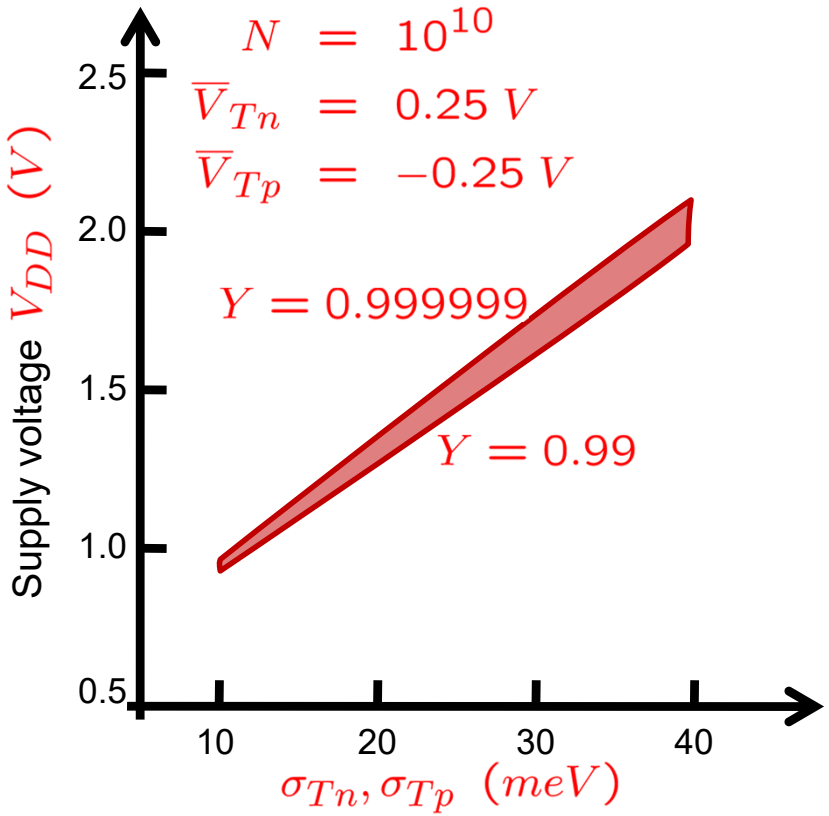
Clock frequency: GHz

Inverter gain: -10

Ensemble size: $N = 10^{10}$

Activity factor: 10%

Model error calculation (Gaussian, threshold)



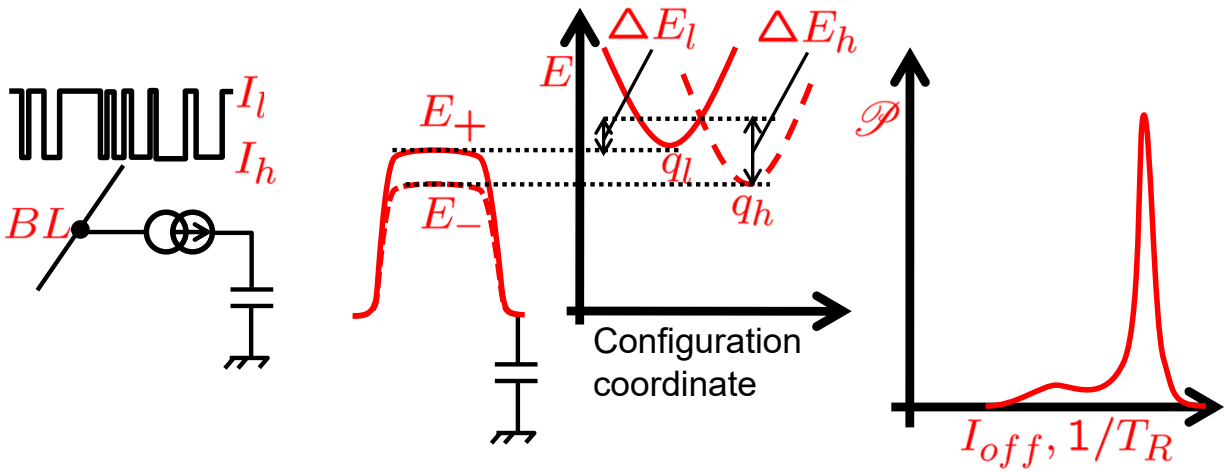
A model square law calculation.

Y : Yield

$$V_{iL} = (1/8)(5V_{Tn} - 3|V_{Tp}| + 3V_{DD},$$
$$V_x = (1/2)(V_{DD} - |V_{Tp}| + V_{Tn})$$
$$V_{DD} = \overline{V}_{Tn} + |V_{tp}^0| + \sigma_{Tn} 5\sqrt{2}\,\mathrm{erf}^{-1}\left[1 - \frac{2(1-Y)}{N}\right]$$

Nominal thresholds 0.25 V for n and the negative for p

Random walk: Variable retention (Poisson)



$$\frac{1}{\tau_h} = \frac{1}{\tau}\exp\left(-\frac{\Delta E_h}{k_BT}\right) \quad \text{and}$$
$$\frac{1}{\tau_l} = \frac{1}{\tau}\exp\left(-\frac{\Delta E_l}{k_BT}\right)$$

$$\Gamma_e = \exp\left(\frac{E_{pf}}{k_BT}\right) + \int_{E_{pf}/k_BT}^{E_G-E_T} \exp\left\{z - \alpha z^{3/2}\left[1 - \left(\frac{E_{pf}}{zk_BT}\right)^{5/3}\right]\right\} dz \quad \text{and}$$

$$\Gamma_c = 1 + \int_0^{E_T} \exp\left\{z - \alpha\left[\left(\frac{E_G}{k_BT}\right)^{3/2} - \left(\frac{E_G}{k_BT} - z\right)^{3/2}\right]\right\} dz$$

$$E_{pf} = (q\mathcal{E}/\pi\epsilon)^{1/2}$$

$$\alpha = 4[2m^*(k_BT)^3]^{1/2}/3q\hbar\mathcal{E}$$

$$I_{HSR} \propto \frac{n_i\Omega}{\tau_l \exp\left[(E_i - E_T)/k_BT\right]} = \frac{1}{\tau}\exp\left[-\frac{(E_G/2) + E_i - E_T + E_l}{k_BT}\right] \approx \frac{1}{\tau}\exp\left(-\frac{E_l - E_T}{k_BT}\right)$$

1 % probability in 10 GB scale memory is 10^8 defective cells.
Needs a power of QV_B/T_R to compensate lower retention cells.
x100 refresh rate

Shannon implications for electric signal flow

Channel capacity in a Gaussian channel: $C = B \log_2 \left(1 + \frac{S}{N}\right)$

$$
\begin{aligned}
B &= 100\, GHz \text{ for } 5\, GHz \text{ clock} \\
C &\approx 2B \\
\therefore\ S/N &= 3
\end{aligned}
$$

$$
\begin{aligned}
&10\% \text{ duty cycle} \\
U &= 14 k_B T \text{ for LVDS}
\end{aligned}
$$

This is the lower limit in energy for deterministic signal flow under idealized message redundancy.
Actuals, without redundancy, will have to be higher.

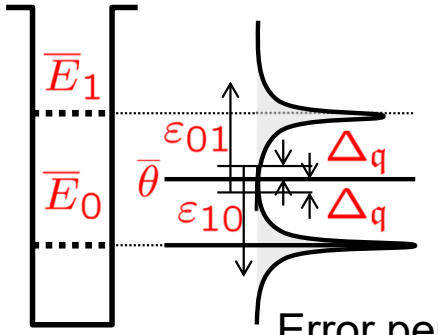
These are all examples of limits placed by determinism in the classical and so enormously successful semiconductor world of logic and memory and traditional computation.

Now a look at some of the ideas to keep the past going.

Quantum wells

Quantum wells are subject to all the material ensemble consequences.
e.g., L must have Gaussian distribution.
Quantum uncertainty will have its i.i.d. contribution.

$$\overline{E}_n = n^2\pi^2\hbar^2/2m^*L^2$$



Error per transition cycle

$$\varepsilon = \frac{1}{2}(\varepsilon_{01}+\varepsilon_{10}) \qquad C = \hbar^2\pi^2/2m^*$$

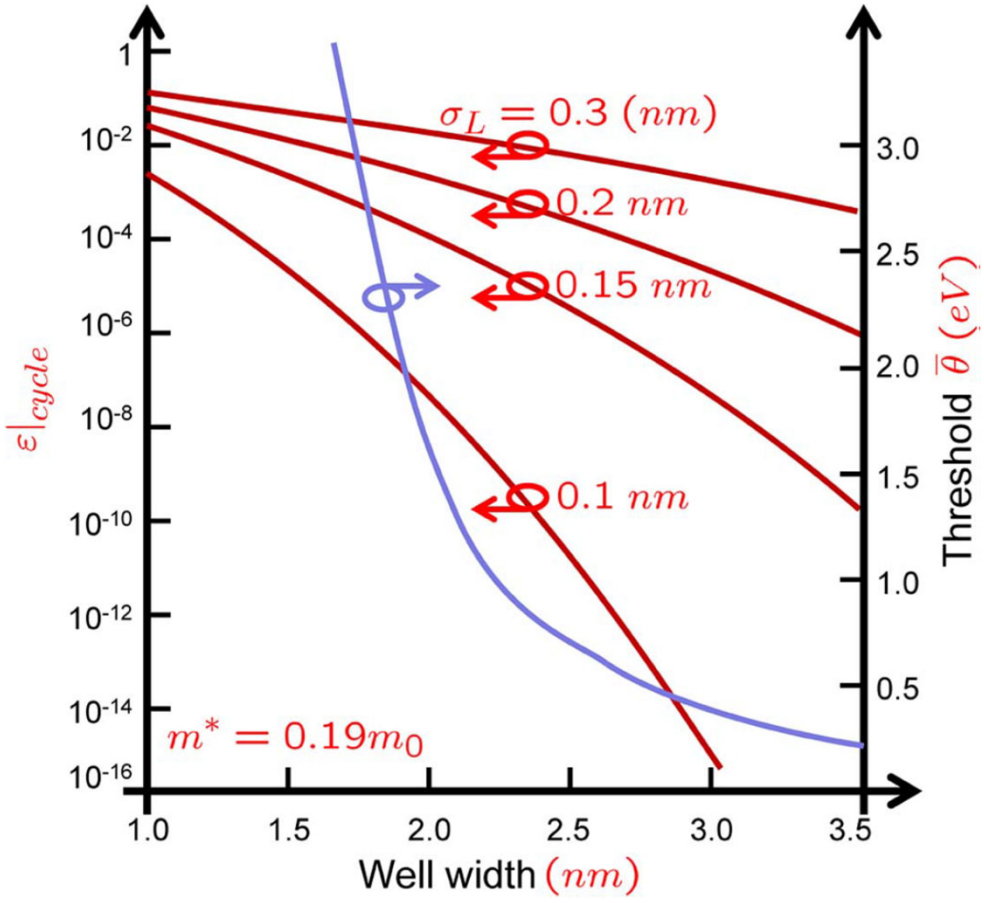
$$= \frac{1}{2}\int_{\overline{\theta}-\Delta_q}^{\infty} \mathfrak{p}(E_1)dE_1 + \frac{1}{2}\int_0^{\overline{\theta}+\Delta_q} \mathfrak{p}(E_2)dE_2$$

$$= -\frac{\sqrt{C}}{4\sqrt{2\pi}\sigma_L}\int_{\overline{\theta}-\Delta_q}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{\sqrt{\frac{C}{E}}-\overline{L}}{\sigma_L}\right)^2\right]\frac{dE}{E^{3/2}}$$

$$-\frac{\sqrt{C}}{2\sqrt{2\pi}\sigma_L}\int_0^{\overline{\theta}+\Delta_q} \exp\left[-\frac{1}{2}\left(\frac{2\sqrt{\frac{C}{E}}-\overline{L}}{\sigma_L}\right)^2\right]\frac{dE}{E^{3/2}}$$

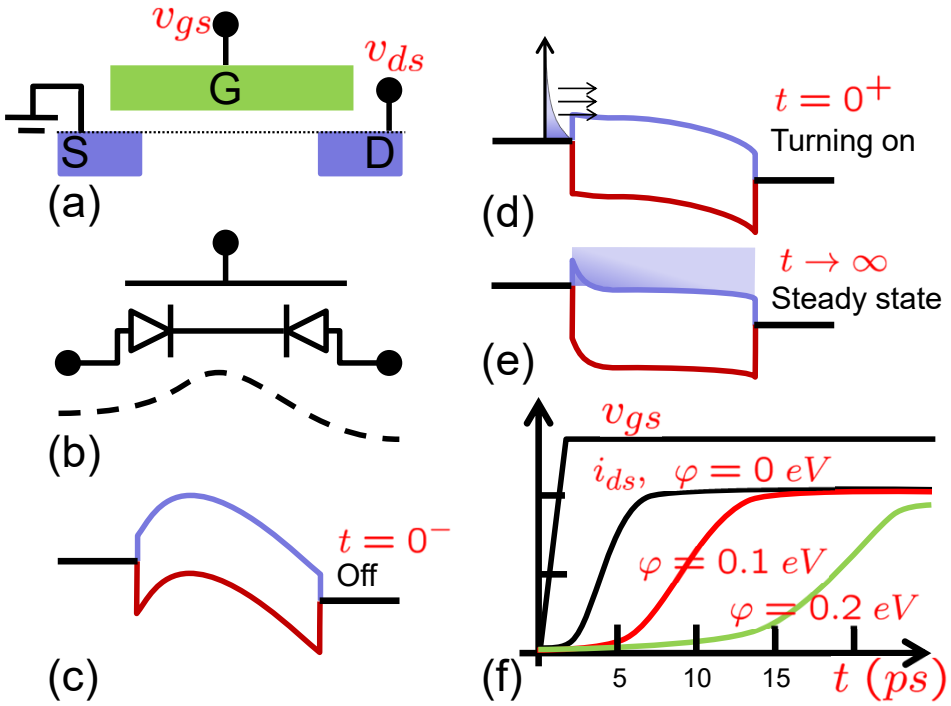
$$= \frac{1}{4}\mathrm{erf}\left(\frac{L_{\overline{\theta}n1}-\overline{L}}{\sigma_L\sqrt{2}}\right) - \frac{1}{4}\mathrm{erf}\left(\frac{\overline{L}}{\sigma_L\sqrt{2}}\right)$$

$$+\frac{1}{4} - \frac{1}{4}\mathrm{erf}\left(\frac{L_{\overline{\theta}n2}-\overline{L}}{\sigma_L\sqrt{2}}\right)$$



Dynamics versus statics: latency consequences

Schottky source and drain



$$J = \frac{2e}{h}\frac{m^*}{(2\pi)\hbar^2}\int_{q\varphi_B}^{\infty} dE_\perp \int_{E_\parallel} dE_\parallel \mathscr{F}\left(\frac{E_\parallel + E_\perp - E_F}{k_BT}\right)$$

$$= \frac{2em^*k_BT}{(2\pi)^2\hbar^3}\int_{q\varphi_B}^{\infty} dE_\perp \ln\left[1 + \exp\left(-\frac{E_\perp - E_F}{k_BT}\right)\right]$$

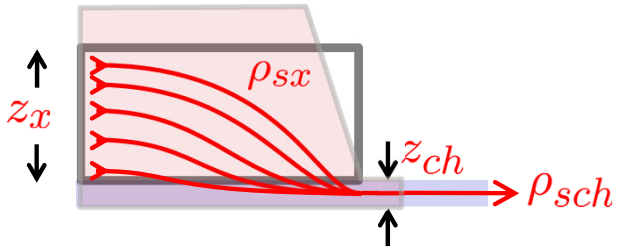
$$\varphi_B \approx 0.2\ eV \Rightarrow \exp(0.2/0.0259) \approx 0.0004$$

$$r = \frac{k_BT}{ei}$$

The absence of a direct control of the electrochemical potential of the contact to the channel junction.

We only have control of the electrochemical potential at the contacts.

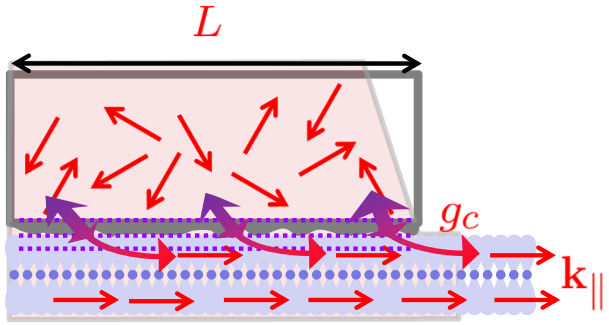
Dimensionality reduction in device structures (Graphene and nanotubes)



$$R_f W = \frac{2}{\pi}\rho_{sx} z_x \ln\left(0.75\frac{z_x}{z_{ch}}\right) - \frac{0.21}{\pi}\rho_{sch} z_{ch}$$

For contact materials such as Pt, say 50 nm thick with a sheet resistance of 200$\Omega/\square$, this resistance is 0.00307 $\Omega.cm$ to inversion layers.

Graphene and nanotubes:



Transmission-line:

$$R_c = \frac{(2\alpha\beta+\beta^2)^{1/2}}{2\beta} R_q \coth\left[(2\alpha\beta+\beta^2)^{1/2} L\right],$$

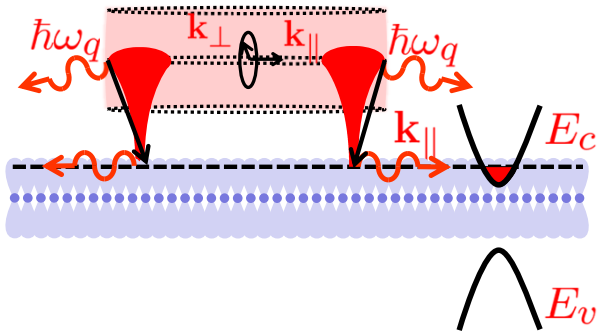
where $\beta = (1/2) g_c R_q$, and $\alpha = \rho_s / R_q$

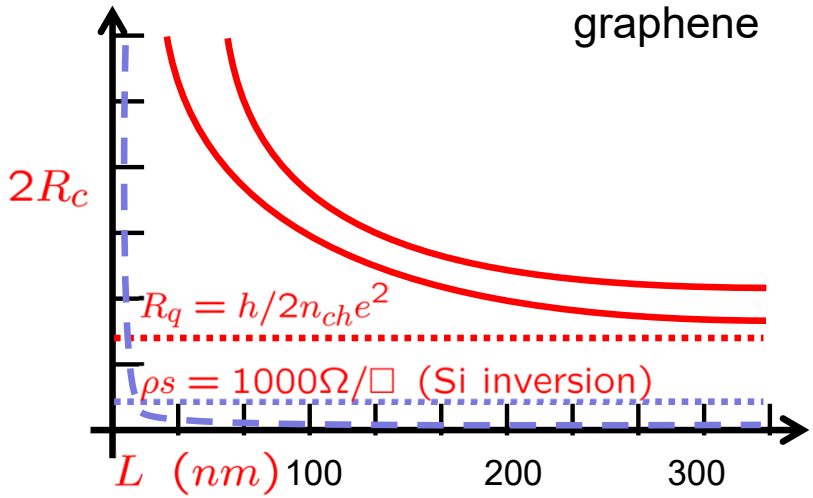
$$\lim_{\alpha\to 0} R_c = \frac{R_q}{2}\coth\left(\frac{1}{2} g_c R_q L\right)$$ Ballistic limit

$$\lim_{\alpha\to\infty} R_c = \sqrt{\rho_{sch}/g_c}\coth\sqrt{g_c/\rho_{sch}}L$$ Diffusive limit

$$\lim_{L\to\infty} R_c = \sqrt{\rho_{sch}/g_c}$$

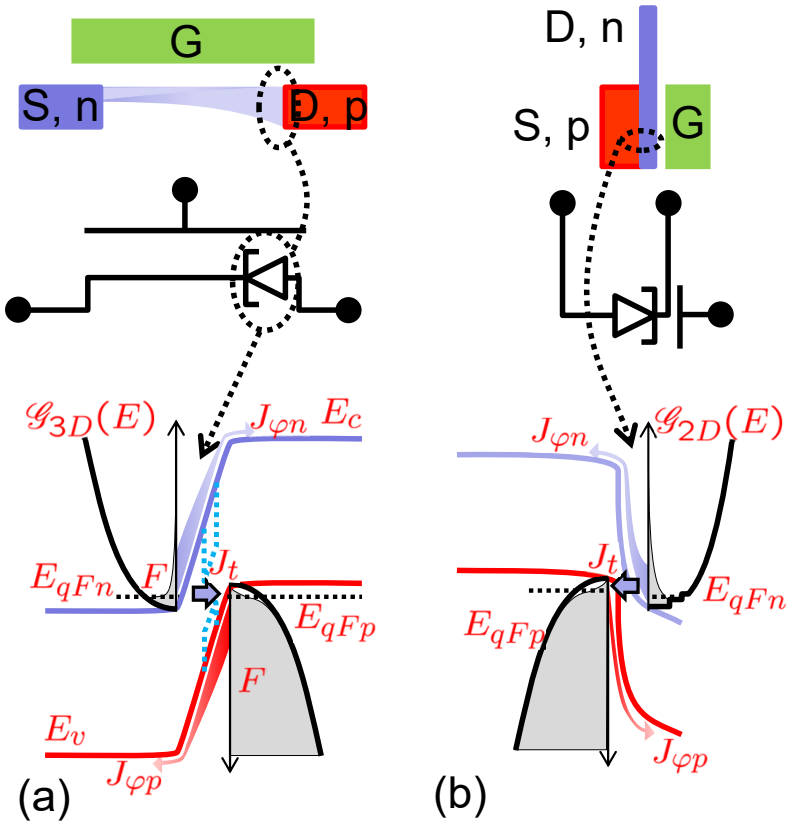
Reciprocal space funneling





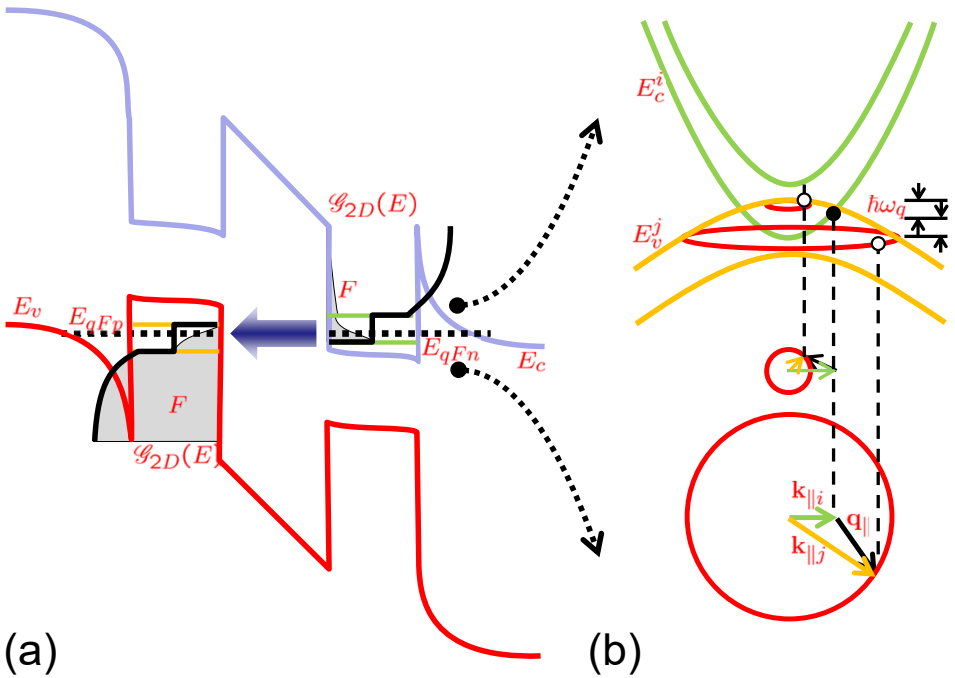
Shortest devices are not limited by gate lengths but by contact lengths.

Subthreshold swing manipulation: Tunneling



Can one restrict state coupling and tune it through the electrochemical modulation from contacts?

A limiting case of tunneling



The argument is based on linewidth and Lorentzian lineshape (i.e., intrinsic quantum).

But statistical fluctuations exist from ensemble There exists an inherent linewidth that is of the order of several 10s of meV

Gaussian fluctuations in step height: Δ

$$\langle \Delta(\mathbf{r})\Delta(\mathbf{r}')\rangle = \Delta^2 \exp(-|\mathbf{r}-\mathbf{r}'|^2/\Lambda^2)$$

$$\propto F_{mn} = \sqrt{(\partial E_m/\partial L)(\partial E_n/\partial L)}$$
$$\propto L^{-3}$$

Cumulative broadening: $\approx 10\text{–}30\ meV$

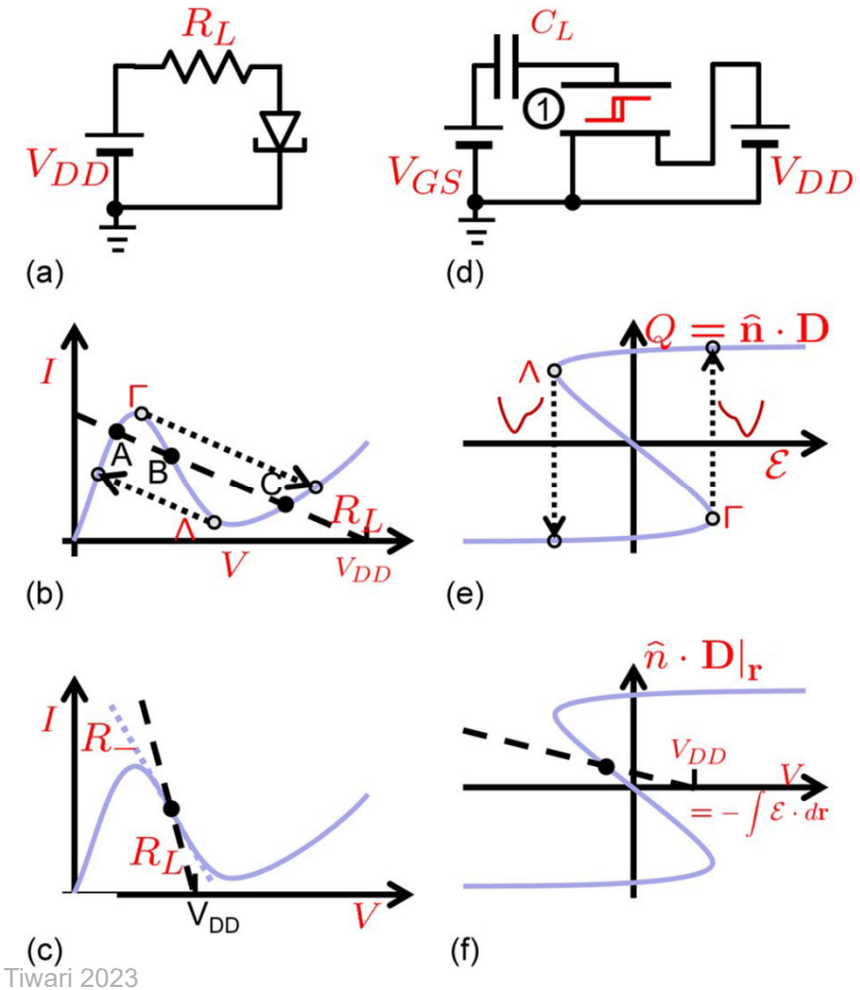
$$J_t = 2|t|^2 \frac{e^2}{\hbar}\mathscr{G}_{2D} V \frac{\Gamma}{(V-V_0)^2+\Gamma^2}$$

$$E_c^i + \frac{\hbar^2 \mathbf{k}_{\|i}^2}{2m_c^*} \pm \hbar\omega_q = E_v^j + \frac{\hbar^2 \mathbf{k}_{\|j}^2}{2m_v^*}$$

$$\mathbf{k}_{\|i} \pm \mathbf{q} = \mathbf{k}_{\|j}$$

Dynamics and their evolution time scale

Tunnel diode Transistor with Ferroelectric

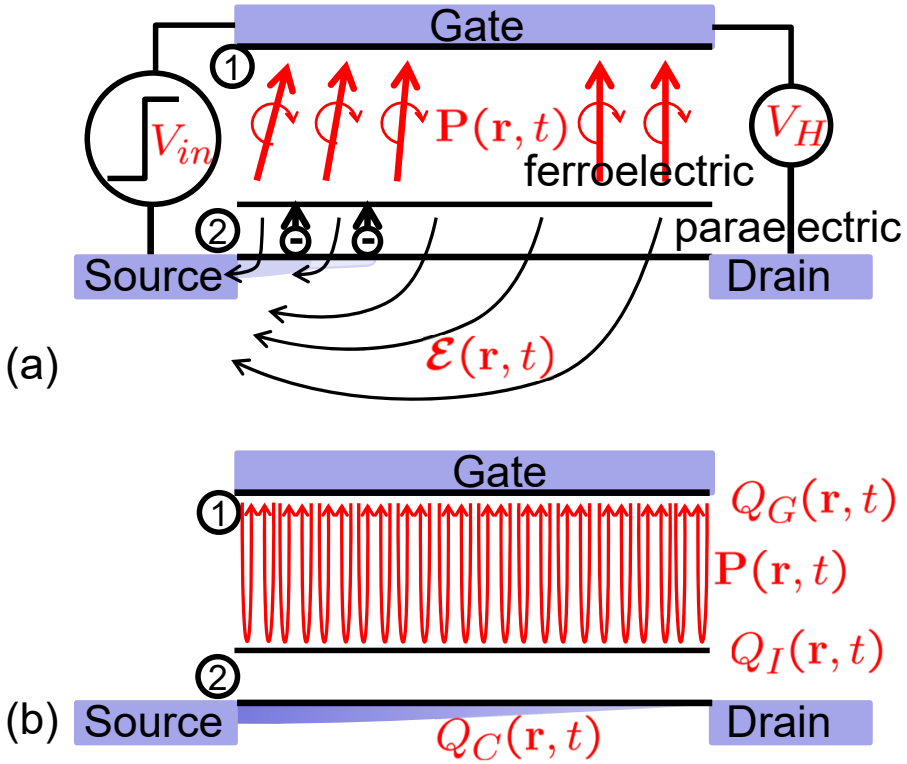


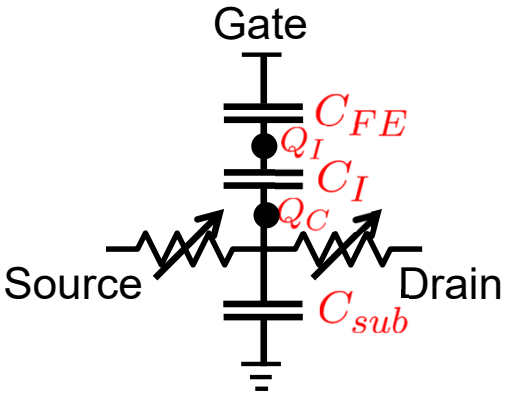
Load lines are a representation of clamping conditions.

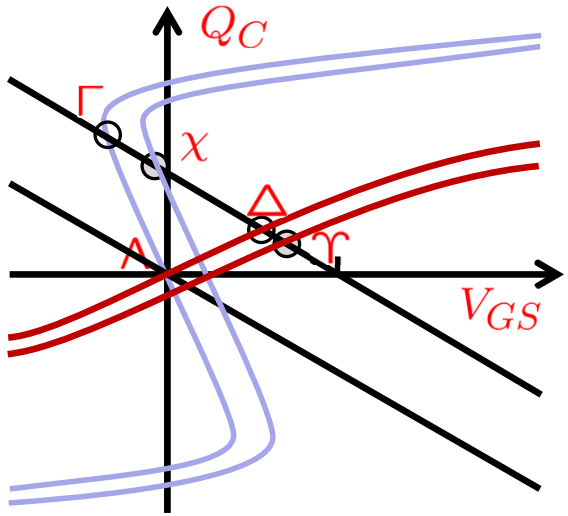
For tunnel diodes, it is all in the electrostatics.

For ferroelectrics, it is within the displacement arising in spontaneous polarization and its evolution. Polarization is structural.

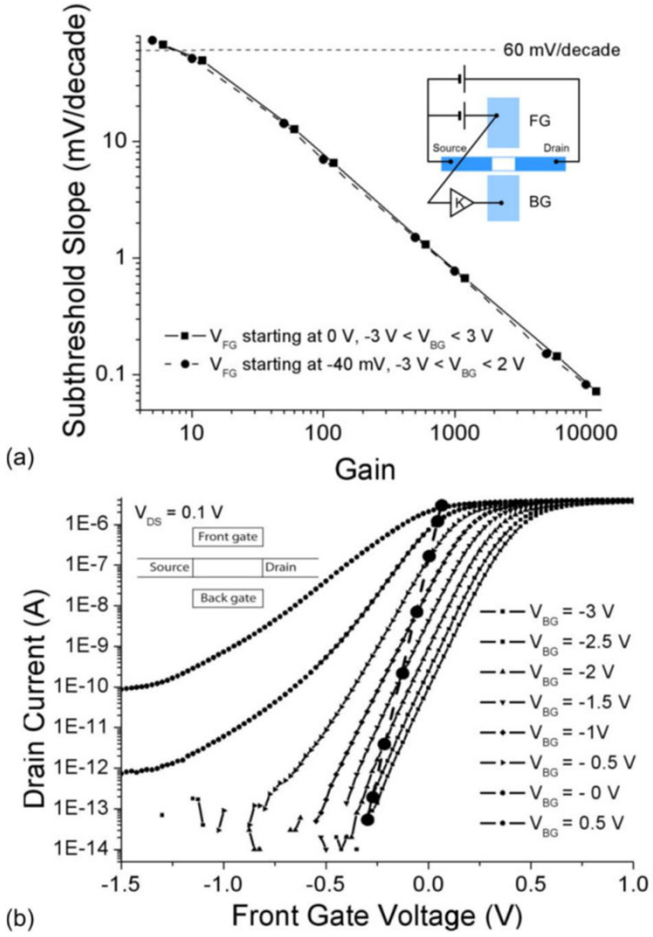
Ferroelectric domain propagation



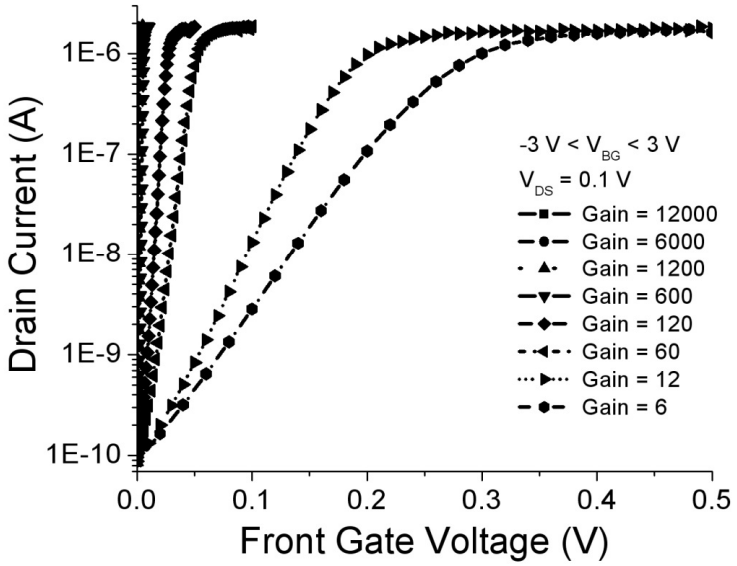




In deterministic classical computing; it is always about energy

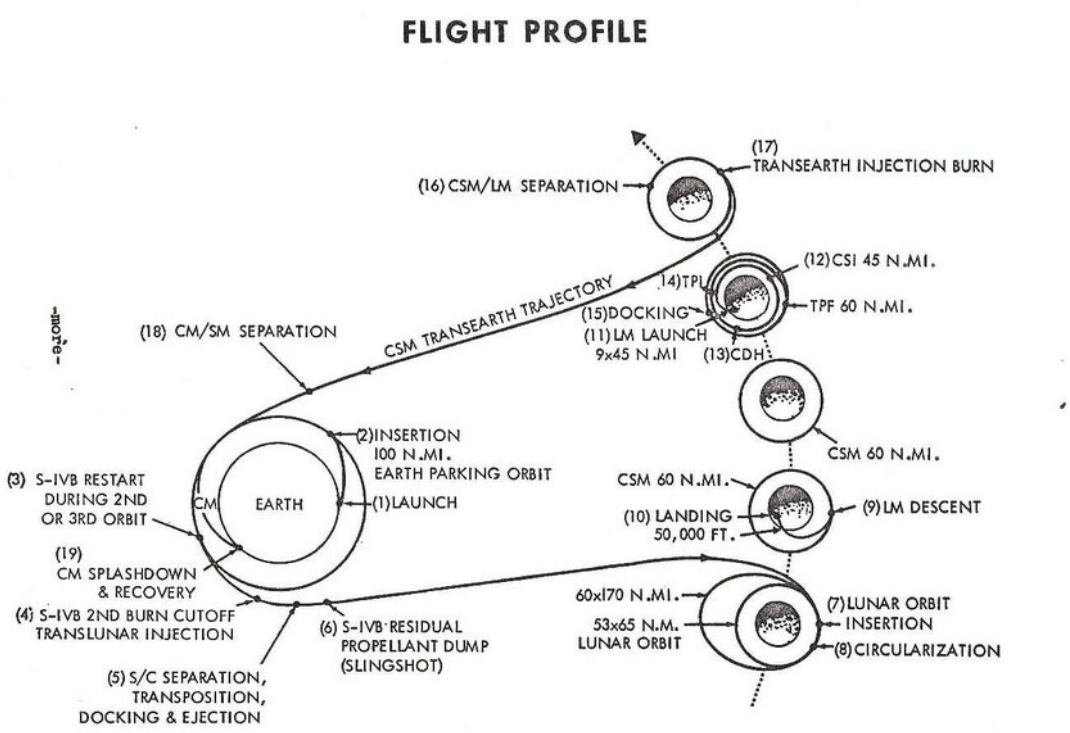


Use of two separate gates with a gain stage in between.



Earth to Moon: Apollo 11

Total Weight of Lunar Module: ~15 tons; Saturn V rockets



~~Hagaromo &~~ Hiten: Low Thrust Transfer

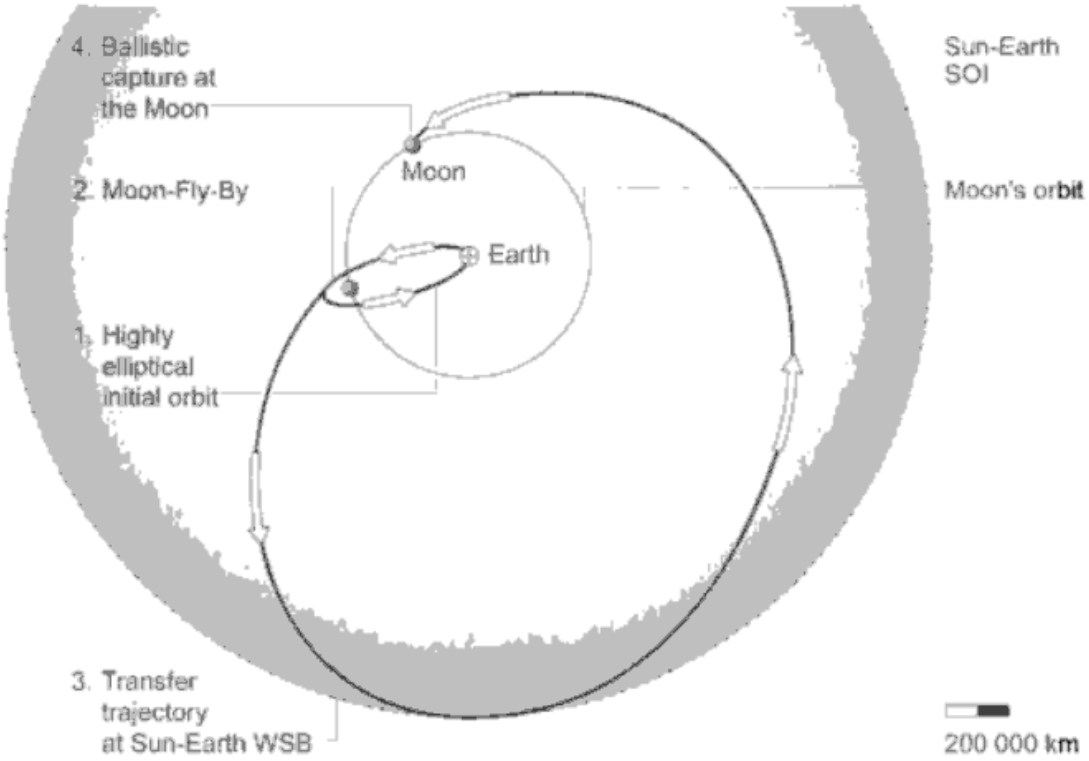
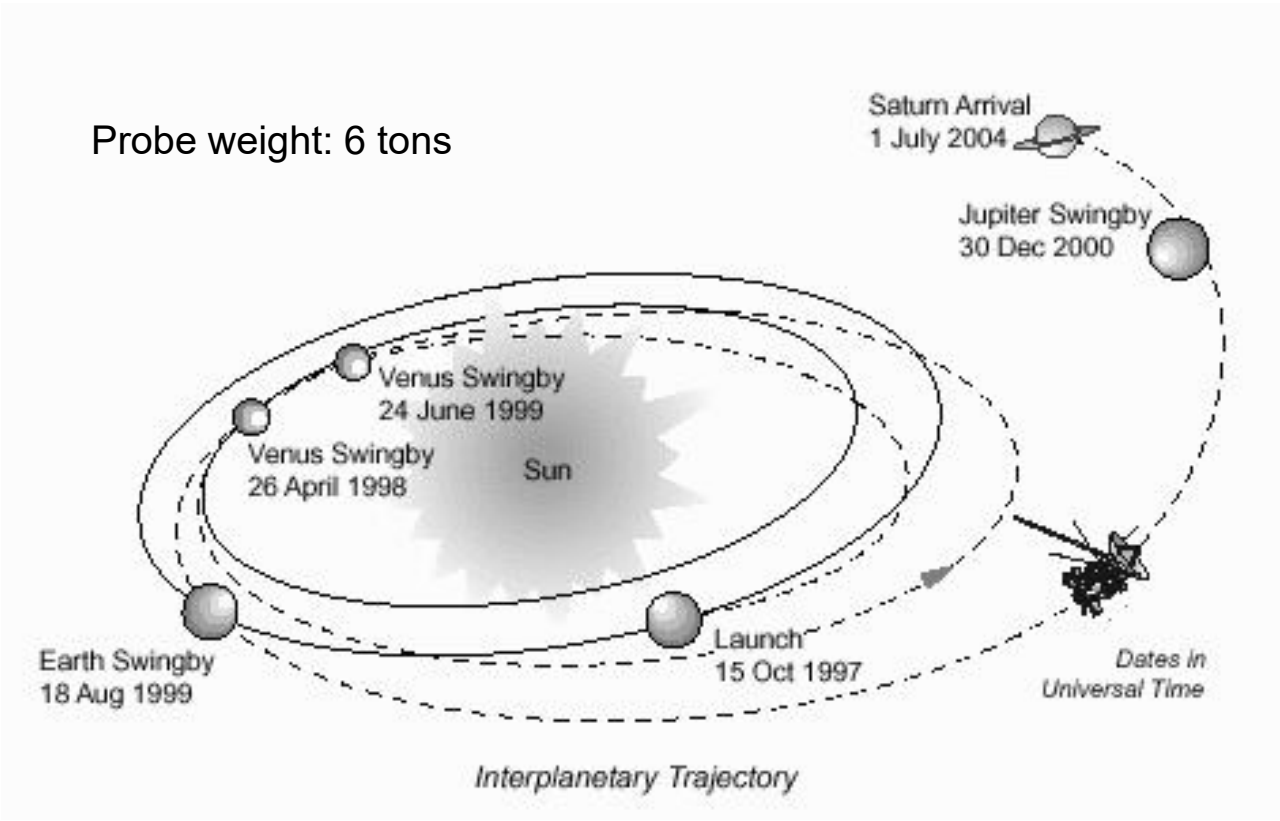


Figure 9.24 The Moon flyby and WSB transfer trajectory of the Japanese Moon probe HITEN.

Cassini: To get to Saturn

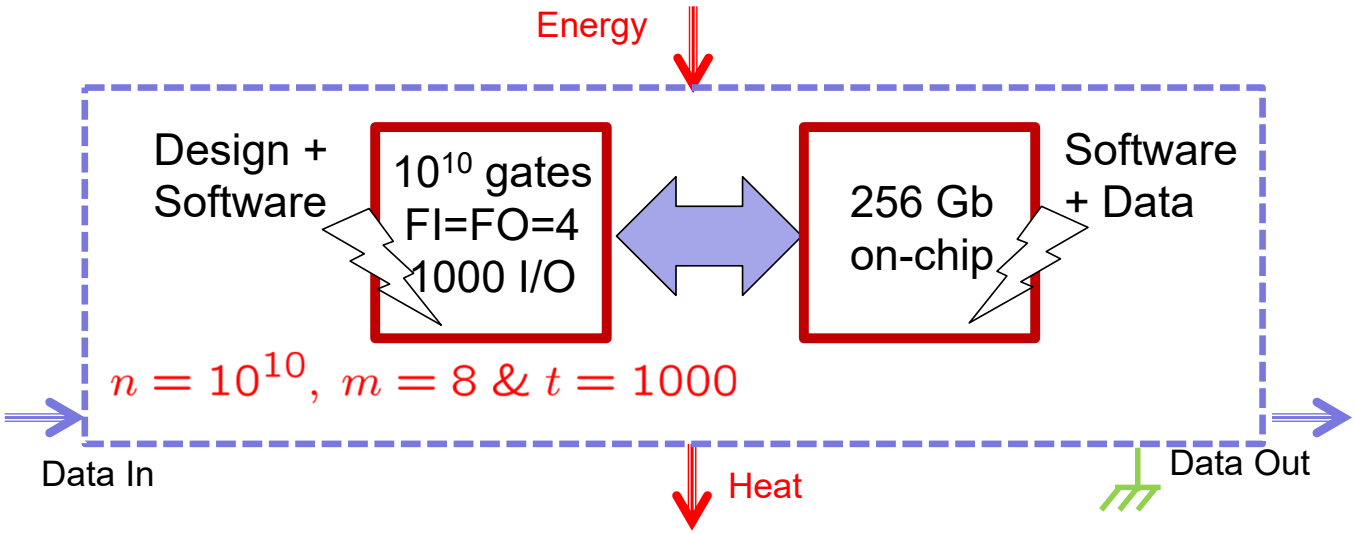
Probe weight: 6 tons



Interplanetary Trajectory

A model of an electronic engine

This is approximately what we carry around in our pocket.



Capacity: $N = nm + t \approx 4 \times 10^{10}$ $\quad N = 256 \times 10^{9}$

Config. volume: $\Omega = N^{N}$ $\quad \Omega = N^{2}$

Shannon Info: $\mathfrak{I} = N \log_2 N \approx 1.5 \times 10^{12}\ b$ $\quad \mathfrak{I} = 2.56 \times 10^{11}\ b$

Negentropy: $\mathfrak{N} = 2.1 \times 10^{-11}\ J/K$ $\quad \mathfrak{N} = 3.5 \times 10^{-12}\ J/K$

Work capacity: $\mathfrak{W} = 7.3\ nW.s$

Structured electronic engine

This structured engine transforms data from implicit to explicit.

e.g., given $g = f(x)$ with f & g known, find $x = f^{-1}(g)$

i.e., make *x* explicit

It does not create data.

The information and organization provided structurally is limited.

e.g., clocks are precisely known.
Have no information content, and all the energy dissipation goes to heat – dumping all of it to ground.

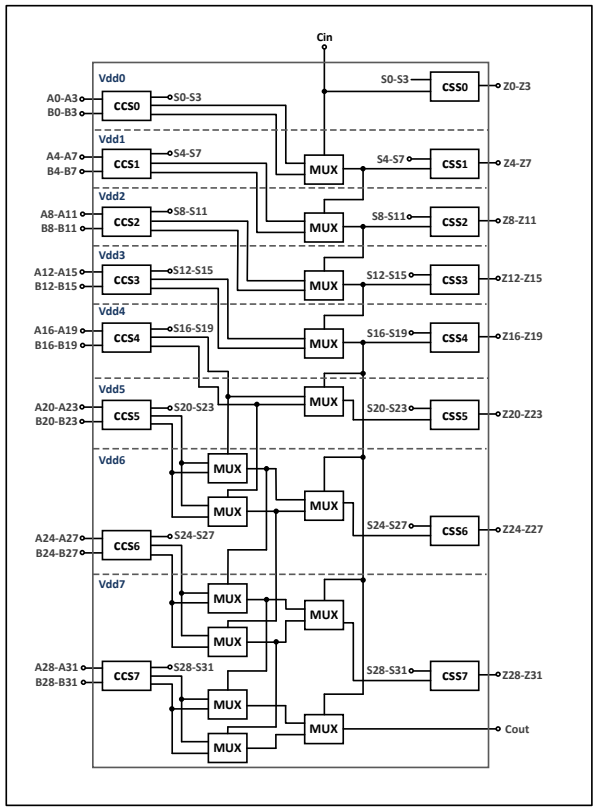
Physical hardware (mechanics of computing) & algorithm (description of how to compute), and the input, sets the upper bound of this data content.

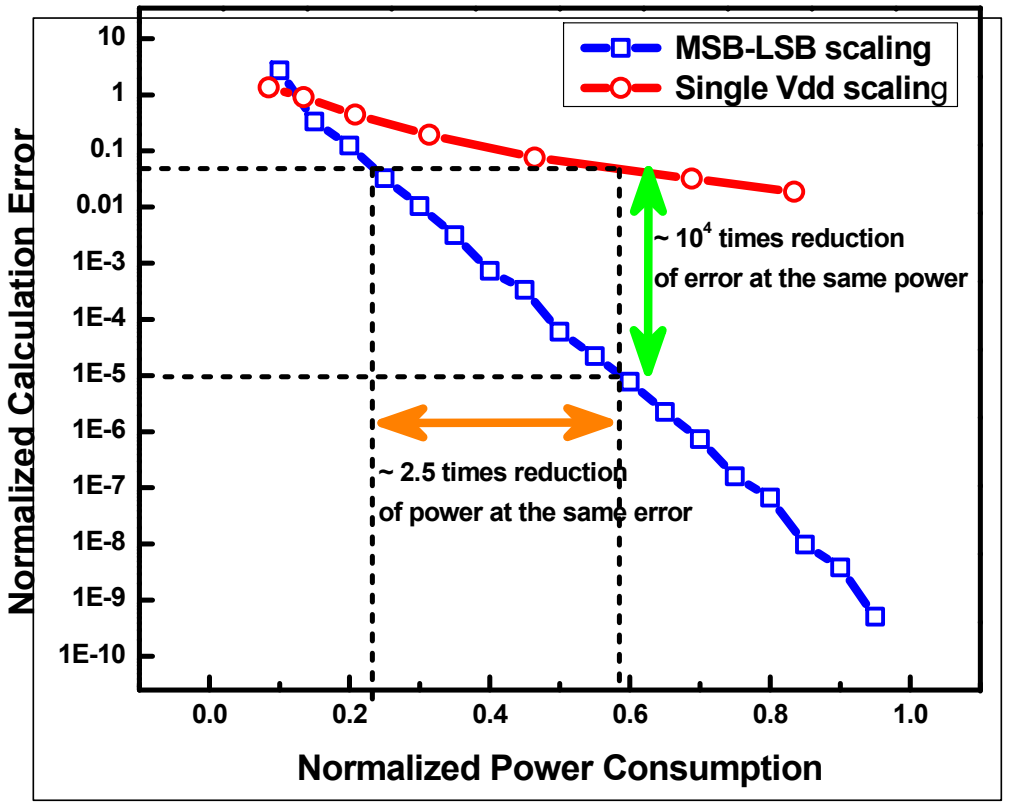
In the thermodynamic sense, the work output of the machine is an improvement in quality of energy – the finesse of information.

Adapting with inexactness

MSB-LSB weighted scaling of supply voltages for *32 b* CCS-CSS adders

Adder Block





Example: Inexact Addition

Floating point sum with *n* *a,b* mantissae

$$A(\equiv a_{n-1}...a_1a_0)+B(\equiv b_{n-1}...b_1b_0) = S(\equiv s_{n-1}...s_1s_0)$$

Consider a sum *S'* within a probability distribution $\mathfrak{p}_\sigma(S^{'} - S)$ centered at *S* of width σ
all expressed normalized – i.e. in units of least precision (*ulp*'s)

If $\sigma \gg 1,$ bits less significant than σ can be ignored and the circuit truncated (i.e. those parts shut off) to make the LSB $\sim \sigma$ *Eliminates wasted power*

$\sigma \sim 1$ is a useful guide mark for error

Consider $FWHM \sim 1$ for

$$\exp\left[-\pi(S-S^{'})^2\right]$$ Gauss

$$\frac{2/\pi}{1+4(S-S^{'})^2}$$ Cauchy-Lorentz

For tails, $|S - S^{'}| \gg 1,$ discreteness can be ignored and what is important is exponential or polynomial tail

Addition

Arithmetic circuit composed of elements: $\alpha = 1, 2, ...M$

For carry-select adders, elements are half adders generating partial sums, and
Kill/Propagate/Generate (KPG) signals.
KPG combine blocks forming carry-propagation tree.
Multiplexers at the output select the partial sums based on carry-propagation tree

For each element α

U_α Energy dissipated per computation using design/voltage scaling

ϵ_α Probability of error

w_α Weight for each element that quantifies mean magnitude of numerical error in the answer caused by incorrectness in the element while all the others are correct

e.g., mux corresponding to bit k of output has the weight 2^k ulp's

Addition

Assume errors in each element uncorrelated

When an element $\alpha \in W \subset 1, 2, ..., M$ is wrong,

the error in answer is: $\langle (S - S^{'})^2 \rangle_W = \sum_{\alpha \in W} w_\alpha^2$

with W as the set of erroneous elements with probability: $\left[\prod_{\alpha \in W} \epsilon_\alpha\right] \left[\prod_{\beta \notin W} (1 - \epsilon_\beta)\right]$

The error in the answer would be within constraints of prob distribution $\mathfrak{p}_\sigma\left((S^{'} - S)^2\right)$ so long as

$$\left[\prod_{\alpha \in W} \epsilon_\alpha\right] \left[\prod_{\beta \notin W} (1 - \epsilon_\beta)\right] \leq \frac{1}{g_W} \mathfrak{p}_\sigma \left(\sum_{\alpha \in W} w_\alpha^2\right) \forall W \subset 1, 2, ..., M$$

Degeneracy – different subsets W that lead to same mean squared error $\sum_{\alpha \in W} w_\alpha^2$

Addition

For Gaussian distribution, the constraint on error is satisfied if

$$\epsilon_\alpha \leq \frac{1}{g_W} \exp\left[-\pi w_\alpha^2\right] \quad \forall \alpha$$

For Lorentzian distribution, the constraint on error is satisfied if

$$\epsilon_\alpha \leq \frac{2/\pi}{g_{w_\alpha}(1+4w_\alpha^2)} \quad \forall \alpha$$

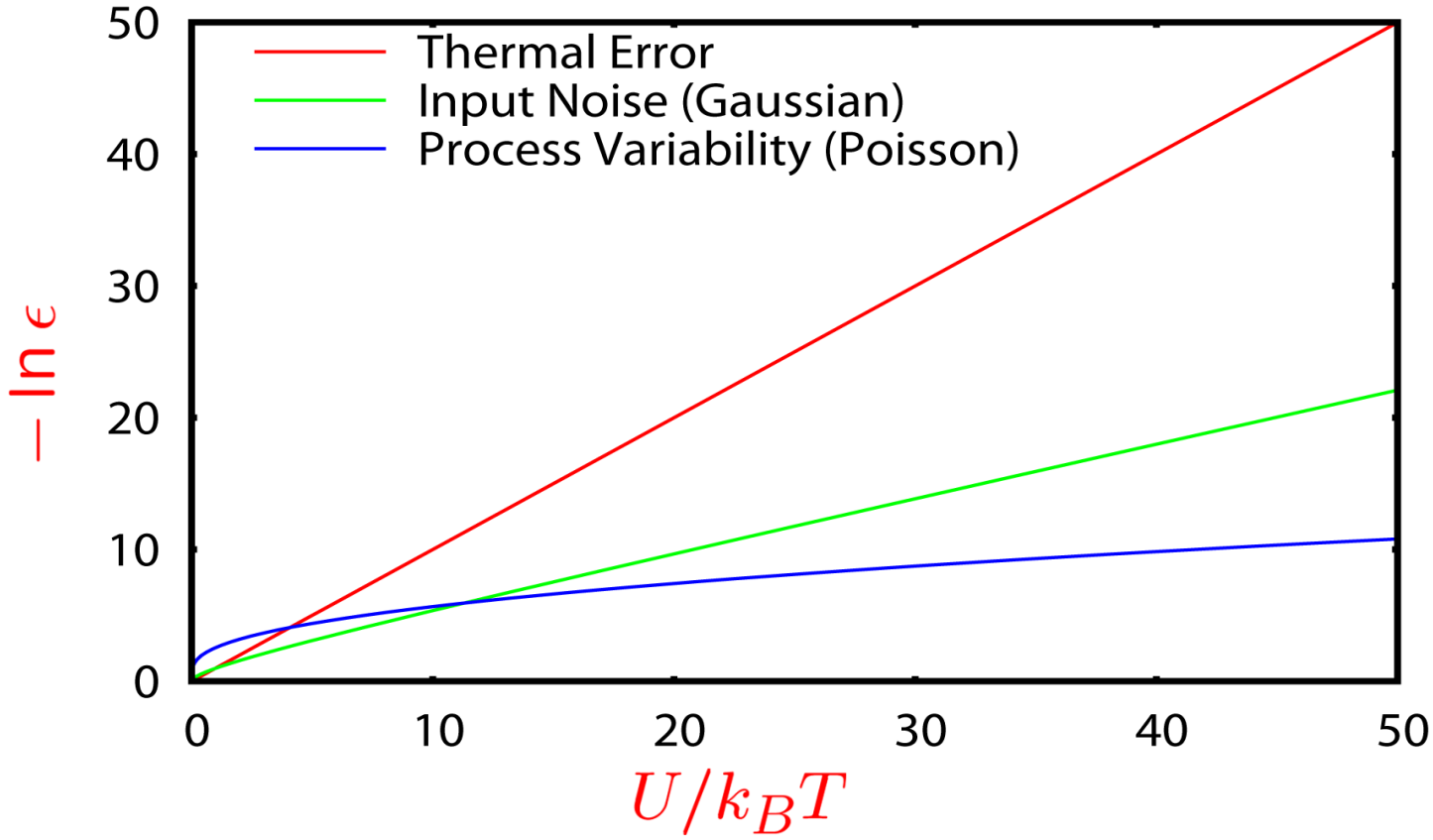
For simple CMOS: $U \propto -\ln \epsilon$

The proportionality constant depends on source of error,

for threshold voltage variations: $\sim CV_{DD}\sigma_{V_T}$

for thermal noise: $\sim k_B T$

Power-Error Trade-Off



By approximate transistor count, half adders are twice as expensive as KPG, Mux and we can estimate energy given the error rate – energy relationship

Adders

Simple case of *carry tree linear chain*: an inefficient ripple-carry adder in carry-select:

One adder, one KPG combine, 1 mux per bit (excluding MSB and LSB end)

Elements at bit level *k* each have a weight 2^k => degeneracy, $g_{w_\alpha} = 3$

Energy dissipated per computation for entire *n*-bit adder :

Gaussian $U \propto -4\sum_{k=0}^{n-1} \ln\left(\frac{1}{3}\exp(-2^{2k}\pi)\right) = 4n\ln 3 + \frac{4\pi}{3}\left(2^{2n}-1\right)$

Lorentzian $U \propto -4\sum_{k=0}^{n-1} \ln\left\{\frac{2/\pi}{3[1+2^2(k+1)]}\right\} \approx 4n\ln\left(\frac{3\pi}{2}\right) + 4n(n+1)\ln 2$

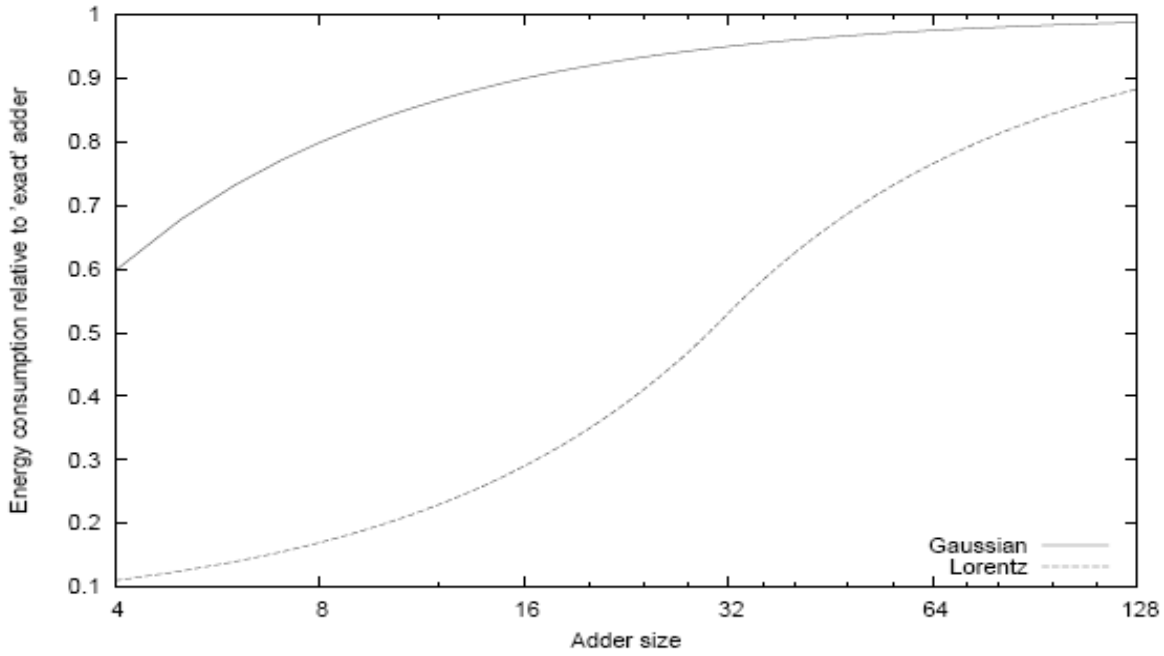
$U \propto 4n\ln\left(\frac{1}{\epsilon}\right)$ (flat error everywhere, i.e., exact arithmetic)

The flat error, exact limit (an estimate of lowest ϵ) can be used to estimate where approximate adders can be more efficient

Example: Ripple Carry Adders

Constraint: no block in adder has an error $< \epsilon_0$,

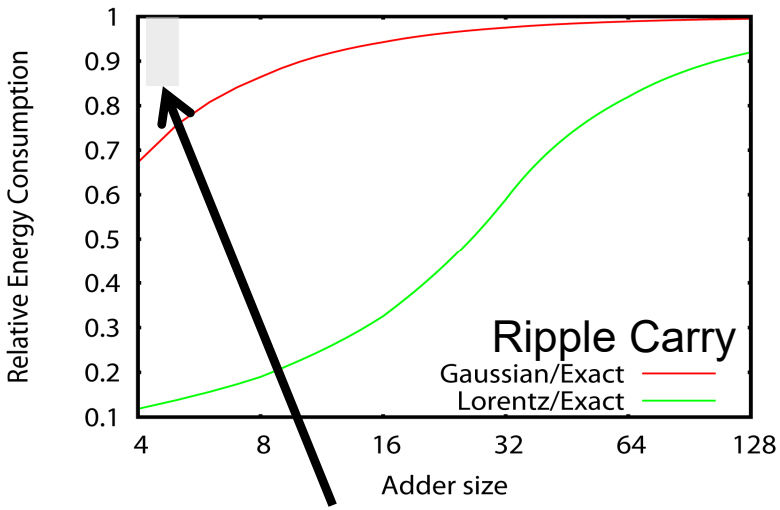
i.e., together with following the distribution $\mathfrak{p}_\sigma(S' - S)$, the answer is random with a miniscule probability $n\epsilon_0$



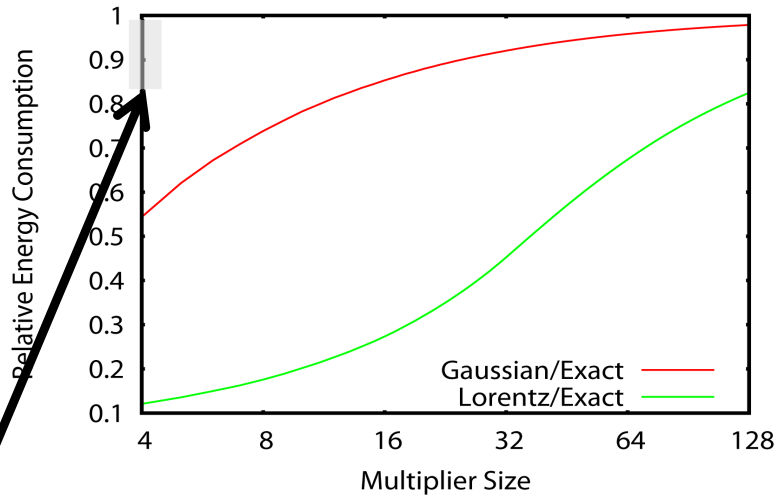
More Useful Cases in Arithmetic Operations

A more efficient adder: Kogge-Stone

KPG-combine block of weight 2^k is $\log_2 k$



Modified Booth Multiplier



Region of maximum power savings

Thank you for hearing me take an old art to the viewing of the modern semiconductor world.

To date, much of the current digital semiconductor effort is all directed to deterministic, error-free implementations.
This extracts a large energy cost.

There are places where one can relax this error-free constraint, as we now do with AI/ML.
This is using deterministic tools for non-deterministic calculations.

There is another step of becoming entirely non-deterministic that may perhaps be useful.
It may also apply to the AI/ML world.
This I will explore in the next two talks.

I thank students over the years, insightful colleagues and my teachers who have influenced this thinking and collaborated in the pursuits.