4IF EESN

A societic-historical perspective on digital hardware (from pebbles to integrated circuits)

en bon français: **du digital au numérique**

Florent de Dinechin

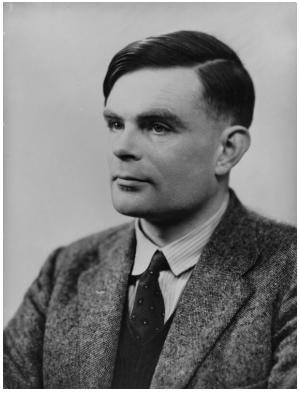
citi lab — INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON — Inria — UNIVERSITÉ DE LYON

Outline

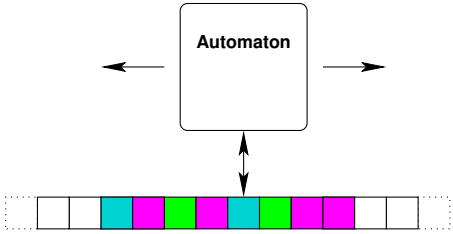
Transition: the war of the programming models

Alan Turing and John von Neumann

... *the twin gods of the computing pantheon* (A. C. Clarke in 2001: A Space Odyssey)

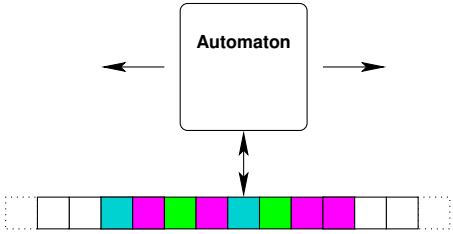


The Turing Machine: a good idea, but a commercial failure



- Defines universality
- Finite automaton infinite memory
- local access to the memory

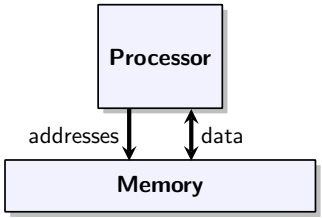
The Turing Machine: a good idea, but a commercial failure

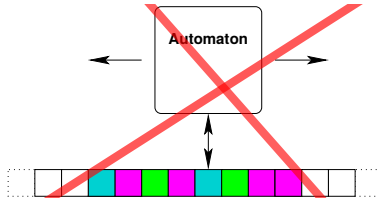


- Defines universality
- Finite automaton infinite memory
- local access to the memory

Kahan: the fast drives out the slow even if the fast is wrong

The von Neumann Machine that ruled

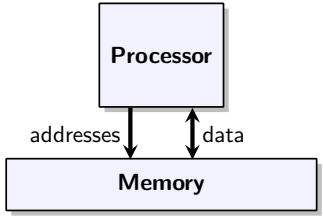


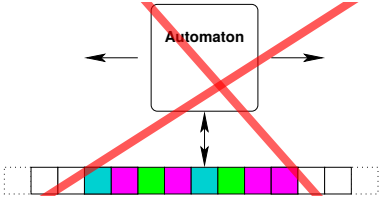


- Universal as well
- Same finite automaton (processor), same infinite[1] memory
- Turing-killer feature: random access to the memory

[1] From there on, infinite means: some power of two, e.g. 2^{32} so large that I can't count that far.

The von Neumann Machine that ruled

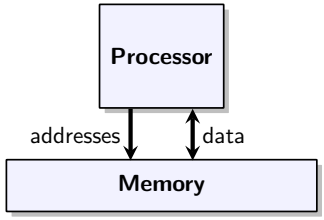


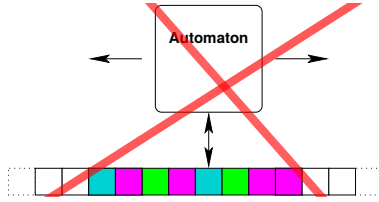


- Universal as well
- Same finite automaton (processor), same infinite[1] memory
- Turing-killer feature: random access to the memory
  - so much more efficient (no tape rewinding)
  - so much easier to exploit (program here, data there, etc)

[1] From there on, infinite means: some power of two, e.g. 2^{32} so large that I can't count that far.

The von Neumann Machine that ruled





- Universal as well
- Same finite automaton (processor), same infinite[1] memory
- Turing-killer feature: random access to the memory
  - so much more efficient (no tape rewinding)
  - so much easier to exploit (program here, data there, etc)

[1] From there on, infinite means: some power of two, e.g. 2^{32} so large that I can't count that far.

When reality kicks back

A law of nature

You can't move data
faster than the speed of light

If the memory is infinite,
some bits will be physically distant
and will therefore be accessed slowly.

When reality kicks back

A law of nature

You can't move data
faster than the speed of light
If the memory is infinite,
some bits will be physically distant
and will therefore be accessed slowly.

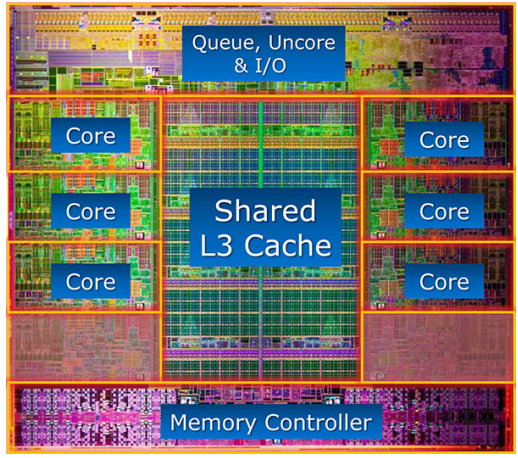
Random access in constant (fast) time
is but a dream
and the von Neumann model
is but a model...

When reality kicks back

A law of nature

You can't move data
faster than the speed of light
If the memory is infinite,
some bits will be physically distant
and will therefore be accessed slowly.

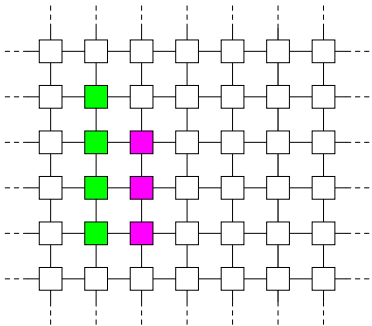
Random access in constant (fast) time
is but a dream
and the von Neumann model
is but a model...



One half of your processor is there to keep this dream alive!

Meanwhile, von Neumann had a better idea

Cellular automaton: a spatial/parallel version of Turing machine.
(most famous instance: Conway's **Game of Life**)



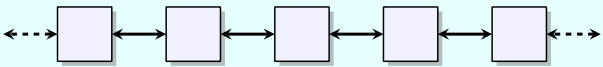
- an infinite number of automata
- all working in parallel
- with next-neighbour (local) communications
- univeral all the same
  (youtube "game of life in game of life")

Let us build this! (with a little help of Moore's law)

Yes but... next-neighbour communications suck!

The firing squad synchronization problem

Synchronize n cells, using next-neighbour communication only

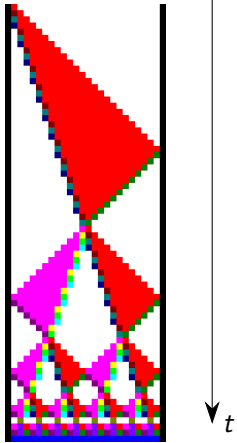


- right: $3n$ steps, using 15 states
- best known: $2n - 2$ steps, 6 states

Jacques Mazoyer.

A six-state minimal time solution to the firing squad synchronization problem. *Theoretical Computer Science*, 50(2):183–238, 1987.

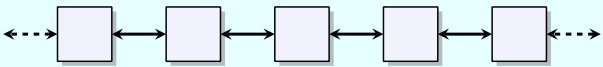
2D version: youtube "game of life in game of life"



Yes but... next-neighbour communications suck!

The firing squad synchronization problem

Synchronize n cells, using next-neighbour communication only

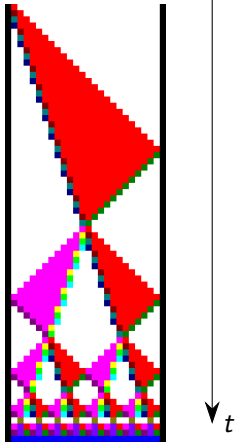


- right: $3n$ steps, using 15 states
- best known: $2n - 2$ steps, 6 states

Jacques Mazoyer.

A six-state minimal time solution to the firing squad synchronization problem. *Theoretical Computer Science*, 50(2):183–238, 1987.

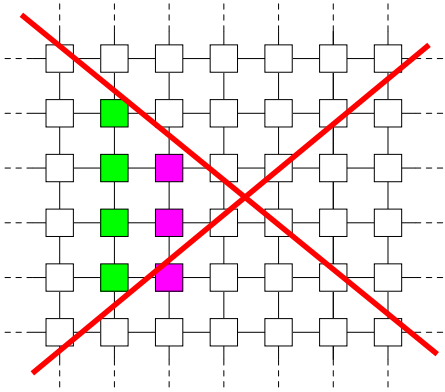
2D version: youtube "game of life in game of life"

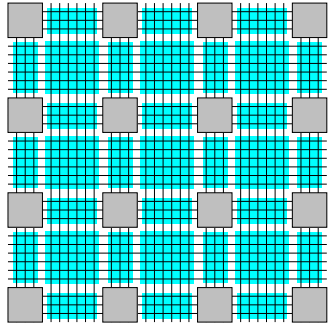


... in real life the captain simply shouts "Fire!" (a global communication)

Field-Programmable Gate Arrays

FPGAs are to 2D cellula automata what von Neuman machines are to Turing machines:
something useful in practice.





Prehistory: who controls numbers controls the world

EESN pour Numérique: l'invention des nombres

Un survol approximatif et incompétent de l'histoire des nombres pour illustrer deux idées :

- des inventions motivées par des Enjeux Sociétaux™®;
- des inventions contraintes par les limitations de la technologie.

Neolithic

Joint invention of

- the unary representation of integers
- the non-volatile memory (*calculus* == pebble)
- the bijection / set idempotency

Enjeu sociétaux:

- compter les moutons
  (alors qu'on ne sait pas encore compter)
- du troc au commerce
  - premiers contrats:
    des cailloux enfermés dans une coquille de terre cuite

Antiquity

alphabetical systems
(Roma, Egypt, Greece, China)

position systems
(Babylonia, India, Maya)

Alphabetical systems

- unary, but several symbols/tokens.
  - coins and bills
  - Roman numerals
- addition and subtraction easy as in unary
  - low-tech addition device: your purse

Can you tell what Societic Challenge is addressed here?

Alphabetical systems

- unary, but several symbols/tokens.
  - coins and bills
  - Roman numerals
- addition and subtraction easy as in unary
  - low-tech addition device: your purse

Can you tell what Societic Challenge is addressed here?

D'après un haut relief de Saqqara en Egypte. Bec dans le vent, les oiseaux indiquent l'ordre de lecture, ici de droite à gauche. Les autres hiéroglyphes comptent les tributs payés à Pharaon après une campagne victorieuse. Chaque signe vaut : un pour la barre, 10 pour le fer à cheval, 100 pour le serpent, 1000 pour le lotus, 10 000 pour l'obélisque et 100 000 pour la salamandre.

Jean Vuillemin, les langages numériques (dispo sur le web)

Position systems

Our decimal system is an example of position system:

- The *position* i of a digit gives its *weight* 10^i
- Example: $789 = 7 \cdot 10^2 + 8 \cdot 10^1 + 9 \cdot 10^0$
- With 3 decimal digits we represent numbers from 000 to 999 $= 10^3 - 1$
- In general, using n digits, we can represent integers in $[0..10^n - 1]$

First advantage of positions systems

A compact representation of arbitrarily large numbers with a fixed number of symbols.

Exponential economic growth can begin...
But also science.

Position systems

Our decimal system is an example of position system:

- The *position* i of a digit gives its *weight* 10^i
- Example: $789 = 7 \cdot 10^2 + 8 \cdot 10^1 + 9 \cdot 10^0$
- With 3 decimal digits we represent numbers from 000 to 999 $= 10^3 - 1$
- In general, using n digits, we can represent integers in $[0..10^n - 1]$

First advantage of positions systems

A compact representation of arbitrarily large numbers with a fixed number of symbols.

Exponential economic growth can begin...
But also science.

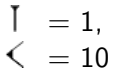
Second and main advantage of positions systems

Algorithms for addition, subtraction, multiplication and division
that scale to arbitrarily large numbers

The Babylonian system

A positional number system in radix 60

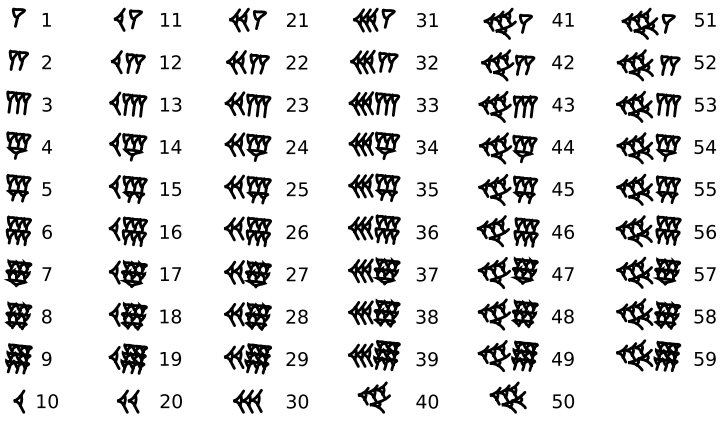
- Radix 60 needs to represent 60 different digits... 60 symbols?
- Each digits represented in a two-symbol alphabetical system:



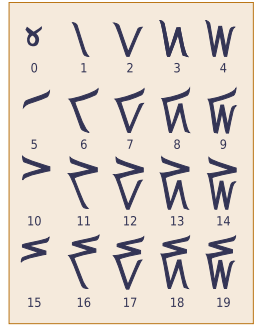
YBC7289 from `http://www.math.ubc.ca/~cass/euclid/ybc/ybc.html`

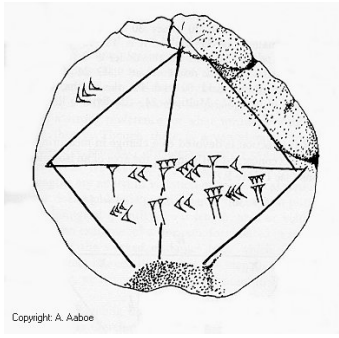
Code superpositions

Babyloniens de jadis



Eskimos de maintenant (Kaktovik)





Une antisèche babylonienne donnant $\sqrt{2}$ et $1/\sqrt{2}$ en base 60.
(on lit 30, autrement dit 1/2, sur le côté du carré,
et sur la diagonale 1:24:51:10 et 42:25:35)

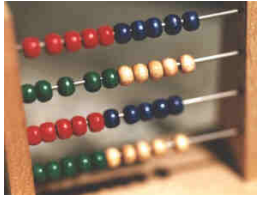
Notation scientifique à exposant implicite

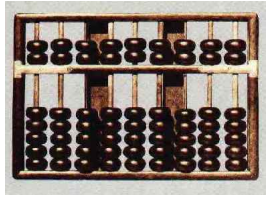
Si je vous dis que j'ai 𒌋𒐗 enfants, cela peut signifier 3/60 enfants, ou 3 enfants, ou 3x60 enfants, ou...
En général, le contexte aide à décider (merci la grande base)

Taking into account the limitations of the technology

- Our brain is able to distinguish at a glance 3 pebbles from 4, but not 7 pebbles from 8...
- On the other hand, it is good at recognizing shapes
- For this reason, Babylonians fixed the shape of the digits

Also reflected in the evolution of the computing device:





Taking into account the limitations of the technology

Digital electronic technology is based on 2-value (Boolean) logic.

Positive integers in the binary position system

- Radix $\beta = 2$
- Two digits 0 and 1
- Bit for "binary digit"
- Larger radices possible by grouping bits :
  e.g. hexadecimal = radix 16, 4-bit digits

Conclusion: la nécessité du calcul

Interro pour voir qui a suivi

A votre avis, les nouveaux systèmes de numération répondaient à un besoin de :

- compter les bulletins de votes pour l'élection du Pharaon
- compter les trimestres ouvrant droit à la retraite pour les tailleurs de pierre
- compter les esclaves, les soldats, et les impôts
- faire avancer la science astronomique

Le numérique accompagne la construction de la civilisation

- Système unaire pour compter les moutons dans de petites communautés
- Systèmes alphabétiques pour compter
  - la richesse dans l'économie réelle
  - mais aussi les soldats (pelotons, compagnies, bataillons, régiments, brigades, divisions)
  - etc.
- Systèmes positionnels pour les calculs plus compliqués :
  - géométrie (un peu pour la science, beaucoup pour le cadastre donc les impôts)
  - lever des impôts
  - astronomie et mathématiques
  - ... et plus tard, physique et économie spéculative

Le numérique accompagne la construction de la civilisation

- Système unaire pour compter les moutons dans de petites communautés
- Systèmes alphabétiques pour compter
  - la richesse dans l'économie réelle
  - mais aussi les soldats (pelotons, compagnies, bataillons, régiments, brigades, divisions)
  - etc.
- Systèmes positionnels pour les calculs plus compliqués :
  - géométrie (un peu pour la science, beaucoup pour le cadastre donc les impôts)
  - lever des impôts
  - astronomie et mathématiques
  - ... et plus tard, physique et économie spéculative

Mais quand est-ce qu'il va nous parler d'ordinateurs ?

Ici bientôt construction d'une transition

Exercice: pour chacun de ces héros, retrouve à quel enjeu sociétal il répond.

- Claude Chappe invente la couverture réseau sans fil
- Ada Lovelace invente le langage de programmation et le multimedia
- Konrad Zuse invente l'ordinateur moderne
- Alan Turing et John von Neumann aussi
- Grace Hopper invente le compilateur
- Margaret Hamilton fonde la profession de hacker, puis pour se faire pardonner invente l'ingéniérie logicielle

Mais sautons tout de suite à l'ère moderne de l'informatique.

History: what kind of law is Moore's Law

Moore's law

From observations in a 1965 paper by Gordon Moore (Intel)

> The number of transistors
> that can be packed on an economically viable chip
> doubles every two years

Moore's law

From observations in a 1965 paper by Gordon Moore (Intel)

> The number of transistors
> that can be packed on an economically viable chip
> doubles every two years

- Mostly a self-fulfilling prophecy
  - If it stops being true, quite a bit of the developed world economy collapses

Moore's law

From observations in a 1965 paper by Gordon Moore (Intel)

> The number of transistors
> that can be packed on an economically viable chip
> doubles every two years

- Mostly a self-fulfilling prophecy
  - If it stops being true, quite a bit of the developed world economy collapses
- Mostly true thanks to the ability to etch *smaller transistors*
  - $\sqrt{2}$ times smaller every other year

Moore's law

From observations in a 1965 paper by Gordon Moore (Intel)

> The number of transistors
> that can be packed on an economically viable chip
> doubles every two years

- Mostly a self-fulfilling prophecy
  - If it stops being true, quite a bit of the developed world economy collapses
- Mostly true thanks to the ability to etch *smaller transistors*
  - $\sqrt{2}$ times smaller every other year
- plus, up to the 70s, improvements in chip area
  - current plateau at 1 cm^2
  - ... for "economically viable"

Moore's law

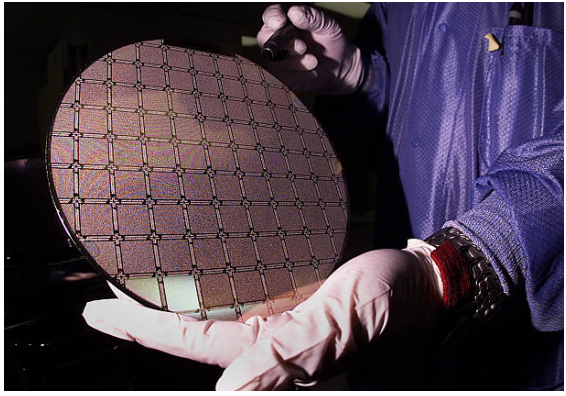
From observations in a 1965 paper by Gordon Moore (Intel)

> The number of transistors
> that can be packed on an economically viable chip
> doubles every two years

- Mostly a self-fulfilling prophecy
  - If it stops being true, quite a bit of the developed world economy collapses
- Mostly true thanks to the ability to etch *smaller transistors*
  - $\sqrt{2}$ times smaller every other year
- plus, up to the 70s, improvements in chip area
  - current plateau at 1 cm^2
  - ... for "economically viable"

From 2004 on: more transistors produced in the world than grains of rice, and cheaper

Economically viable ?



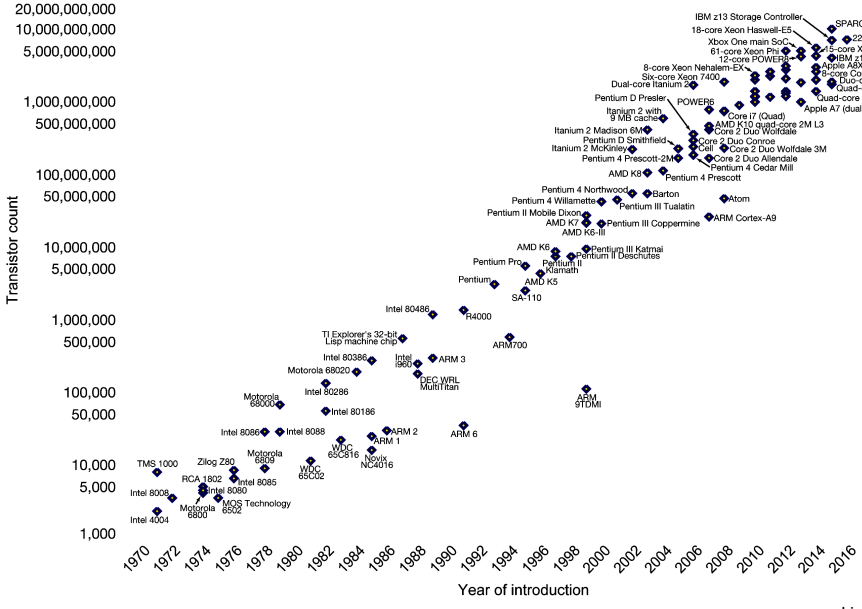
Very expensive and very clean factories

- The wafer size is fixed
  (30 cm since 2002)
- Still, despite the gloves and stuff,
  about 10 defects per wafer

"Yield" of the fabrication process

- 4 chips per wafer:
  very low chance that one works
- 20 chips per wafer: 50% should work
- 1000 chips per wafer: 99% should work

Gamerz GPU: about 5 cm^2 (about 100/wafer); notebook processor about $1cm^2$
Wafer-scale circuits possible *only* if they are designed to be resilient to defects (Cerebras)



Licence CC, Source: Wikipedia

Dennard scaling

From a 1974 paper by Robert Dennard (IBM)

Smaller transistors

- run faster, and
- consume less

so that overall, chip-level dissipated power mostly constant

Dennard scaling

From a 1974 paper by Robert Dennard (IBM)

Smaller transistors

- run faster, and
- consume less

so that overall, chip-level dissipated power mostly constant

Dennard scaling stopped in 2004.

What happened to Dennard scaling?

Smaller transistors

- still could run faster, but
- no longer consume less

so more transistors per chip (Moore) entails more power dissipation per chip.

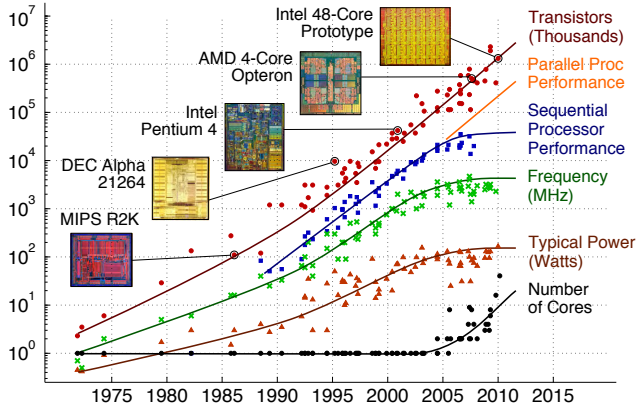
The problem is to move the heat out

Practical power dissipation limit: $\mathbf{100W/cm^2}$
10x your cooking pan, comparable to the rods of a nuclear power plant

Remark: 3D integration helps Moore, but annoys Dennard even more.

The current solution to the end of Dennard scaling

Trend 2: Multicore Performance Scaling



Data partially collected by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond

Nobody asked for multicores !

Life was simpler with single-core programming.

The great depression

- Edward Lee: The Problem With Threads, 2006
- David Patterson: The Trouble With Multicore, 2010

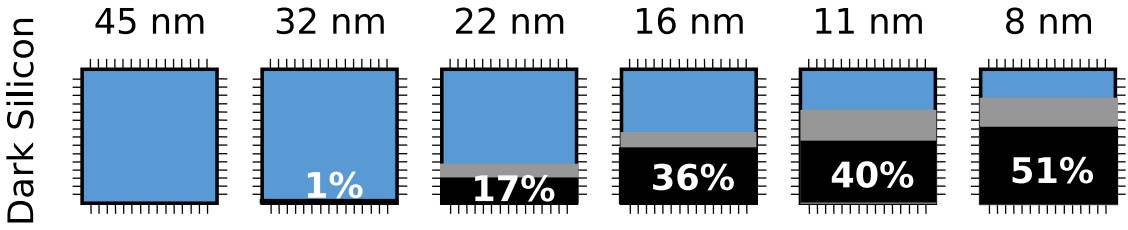
Homework: go read them.

The dark silicon apocalypse

Dark silicon?

In current tech, you can no longer
use 100% of the transistors 100% of the time
without destroying your chip.

"Dark silicon" is the percentage that must be off at a given time



(picture from a 2013 HiPEAC keynote by Doug Burger)

Pleasant times to be an architect

One way out the dark silicon apocalypse (M.B. Taylor, 2012)

Hardware implementations of rare (but useful) operations:

- when used, dramatically reduce the energy per operation (compared to a software implementation that would take many more cycles)
- when unused, serve as radiator for the used parts

Since they are rare, nobody bothered to study them before...

Reality shouldn't constraint our formalisms

Reality shouldn't constraint our formalisms

The end of Moore's law

- Size of an atom?
  The mesh size in silicon crystal is about 0.5nm (1nm=10^{-9}m).

Reality shouldn't constraint our formalisms

The end of Moore's law

- Size of an atom?
  The mesh size in silicon crystal is about 0.5nm (1nm=10^{-9}m).
- Current technology is marketed as 5nm: this corresponds to 10 atoms wide.

Reality shouldn't constraint our formalisms

The end of Moore's law

- Size of an atom?
  The mesh size in silicon crystal is about 0.5nm (1nm=10^{-9}m).
- Current technology is marketed as 5nm: this corresponds to 10 atoms wide.
- Corresponding oxide layer is about two atoms high, and won't get much thinner.

Reality shouldn't constraint our formalisms

The end of Moore's law

- Size of an atom?
  The mesh size in silicon crystal is about 0.5nm (1nm=10^{-9}m).
- Current technology is marketed as 5nm: this corresponds to 10 atoms wide.
- Corresponding oxide layer is about two atoms high, and won't get much thinner.

Remarks: these nanometers used to measure the width of a wire, but it got complicated and it is now pure marketing

Reality shouldn't constraint our formalisms

The end of Moore's law

- Size of an atom?
  The mesh size in silicon crystal is about 0.5nm (1nm=10^{-9}m).
- Current technology is marketed as 5nm: this corresponds to 10 atoms wide.
- Corresponding oxide layer is about two atoms high, and won't get much thinner.

Remarks: these nanometers used to measure the width of a wire, but it got complicated and it is now pure marketing

The end of Dennard scaling

- Corresponding oxide layer is about two atoms high.
  $\longrightarrow$ quantum tunelling $\longrightarrow$ power waste

Reality shouldn't constraint our formalisms

The end of Moore's law

- Size of an atom?
  The mesh size in silicon crystal is about 0.5nm (1nm=10^{-9}m).
- Current technology is marketed as 5nm: this corresponds to 10 atoms wide.
- Corresponding oxide layer is about two atoms high, and won't get much thinner.

Remarks: these nanometers used to measure the width of a wire, but it got complicated and it is now pure marketing

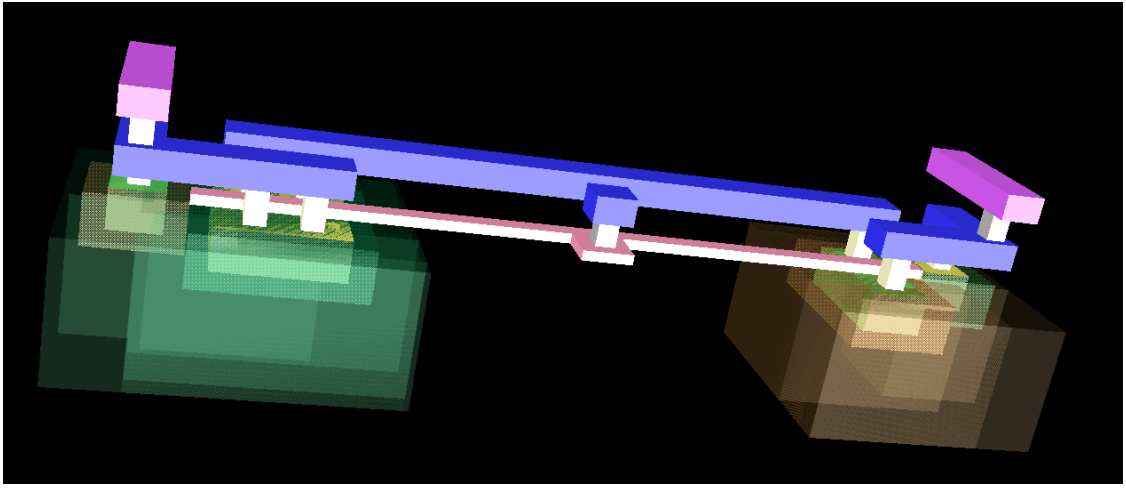
The end of Dennard scaling

- Corresponding oxide layer is about two atoms high.
  $\longrightarrow$ quantum tunelling $\longrightarrow$ power waste
- Transistor threshold voltage got down from 5V to 1V, and won't go much lower

Oxide layer?

The following picture is advertising for the Electric CAD software

http://www.staticfreesoft.com/



Reality shouldn't constraint our formalisms

Other limits

- Speed of light?

Reality shouldn't constraint our formalisms

Other limits

- Speed of light? $3 \cdot 10^8$ m/s.

Reality shouldn't constraint our formalisms

Other limits

- Speed of light? $3 \cdot 10^8$ m/s.
- At the speed of light, a 3GHz signal travels no further than 10 cm in a period
- Homework: cross this with atom size, and get a limit frequency

It's the economy, stupid

The economic cost of a self-fulfilling prophecy

Each new foundry is twice as expensive as the previous one
(or: the cost of a new foundry also follows Moore's law)

- Why?
  - Building billions of reliable objects, each 30 atoms wide
    requires a pretty good vacuum cleaner...
  - Lithographic process used light, then UV, now X rays...
- So foundries are merging to share the costs
  - In 2023, TSMC and Samsung control 70% of the market.
  - ... at some point there will be no competitor left to merge with.
  - (already the case for the manufacturers of foundry equipment, look up ASML)

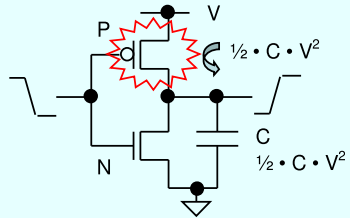
It's the energy, stupid

Back to physics:

Computing consumes energy

- Each bit flipped entails a transfer of electrons from the $\ominus$ to the $\oplus$ through some resistors
- Currently, switching 1 bit costs $10^{-18} J$

(1 attoJoule)



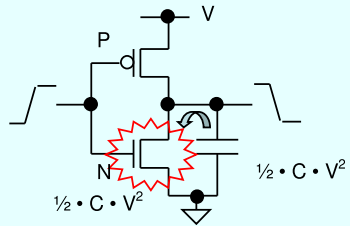


Figure from *Energy per Instruction Trends in Intel Microprocessors* by Grochowski and Annavaramx

It's the energy, stupid (2)

Moving bits consumes energy

- Switching 1 bit costs $10^{-18} J$,
- Moving 1 bit costs $10^{-12} J/cm$ $10^6\times$ more!

(Really the same drawing, but with a larger C)

Remark: there are several hundreds of km of wires inside your processor

It's the energy, stupid (2)

Moving bits consumes energy

- Switching 1 bit costs $10^{-18} J$,
- Moving 1 bit costs $10^{-12} J/cm$ $10^6\times$ more!

(Really the same drawing, but with a larger C)

Remark: there are several hundreds of km of wires inside your processor

Doing nothing consumes energy

These days, roughly 1/3rd of power is leaked (quantum tunelling, etc).

It's the energy, stupid (3): the macro view

Approximate power in 28nm processor (adapted from Bill Dally)

- One 64-bit floating-point Fused Multiply-Add: **50 pJ**
  - This includes switching and moving around inside the FMA
- Access to a 1Kx256-bit on-chip SRAM: **50 pJ**

It's the energy, stupid (3): the macro view

Approximate power in 28nm processor (adapted from Bill Dally)

- One 64-bit floating-point Fused Multiply-Add: **50 pJ**
  - This includes switching and moving around inside the FMA
- Access to a 1Kx256-bit on-chip SRAM: **50 pJ**
- Moving 64 bits 1cm on-chip: **64 pJ**

It's the energy, stupid (3): the macro view

Approximate power in 28nm processor (adapted from Bill Dally)

- One 64-bit floating-point Fused Multiply-Add: **50 pJ**
  - This includes switching and moving around inside the FMA
- Access to a 1Kx256-bit on-chip SRAM: **50 pJ**
- Moving 64 bits 1cm on-chip: **64 pJ**
- Reading 64 bits from external DRAM: **4000 pJ**
  - due to the capacity of macro wires (between chips on your main board)

It's the energy, stupid (3): the macro view

Approximate power in 28nm processor (adapted from Bill Dally)

- One 64-bit floating-point Fused Multiply-Add: **50 pJ**
  - This includes switching and moving around inside the FMA
- Access to a 1Kx256-bit on-chip SRAM: **50 pJ**
- Moving 64 bits 1cm on-chip: **64 pJ**
- Reading 64 bits from external DRAM: **4000 pJ**
  - due to the capacity of macro wires (between chips on your main board)

Some people will tell you...

that doing the same operation in the Cloud (10^4 times further away) saves energy.
Exercise your critical spirit.

(Drew DeVault: "AI/blockchain/bitcoin/climate-disaster-as-a-service")

Hence the current trends in VLSI circuits

Exposed here very well by Christopher Batten:

`https://web.csl.cornell.edu/engrg1060/handouts/engrg1060-ece-lecture.pdf`

Preparing for post-history

Gordon Moore was also an engineer

... and engineers, unlike most economists, know one thing about exponentials:

Gordon Moore in 2005

"It can't continue forever. The nature of exponentials is that you push them out and eventually disaster happens."

(speaking of Moore's Law, but could apply as well to any economic growth of 5%/year).

Moore passed away just two years ago

The International Technology Roadmap for Semiconductor

The ITRS was the consortium in charge of fulfilling the prophecy.

- hundreds of participants, including application people
- addressing gate and RAM technology, but also factory equipment (vacuum cleaners, UV lamps, etc.)

They self-dissolved in 2016, after publishing their final 5-year roadmap.

Moore passed away just two years ago

The International Technology Roadmap for Semiconductor

The ITRS was the consortium in charge of fulfilling the prophecy.

- hundreds of participants, including application people
- addressing gate and RAM technology, but also factory equipment (vacuum cleaners, UV lamps, etc.)

They self-dissolved in 2016, after publishing their final 5-year roadmap.

- NVIDIA's CEO stated in september 2022 that Moore's law is dead.
- Intel's CEO replied one week later "même pas vrai". And indeed it still crawls a bit:
  - FinFET transistors
  - non-volatile RAMs around the corner, and memristors and ...
  - ... and some day the successor of the transistor

Moore passed away just two years ago

The International Technology Roadmap for Semiconductor

The ITRS was the consortium in charge of fulfilling the prophecy.

- hundreds of participants, including application people
- addressing gate and RAM technology, but also factory equipment (vacuum cleaners, UV lamps, etc.)

They self-dissolved in 2016, after publishing their final 5-year roadmap.

- NVIDIA's CEO stated in september 2022 that Moore's law is dead.
- Intel's CEO replied one week later "même pas vrai". And indeed it still crawls a bit:
  - FinFET transistors
  - non-volatile RAMs around the corner, and memristors and ...
  - ... and some day the successor of the transistor

Some of the industry is probably headed for disaster.

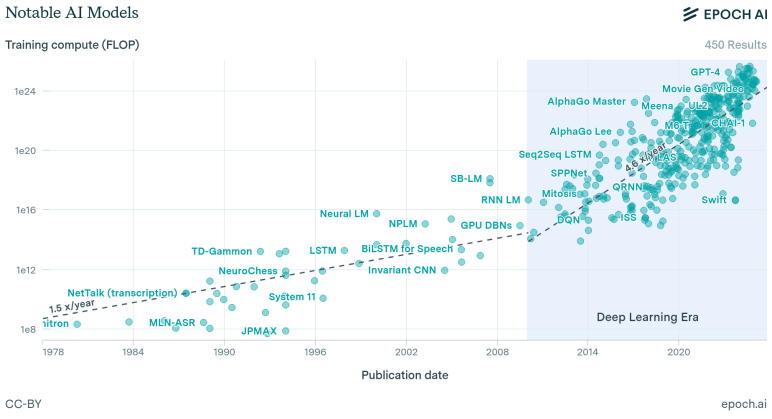
However, there will also be a (very similar, but more serious) *petrol* disaster,
and at least two other, actually serious, disasters: *climate* and *biodiversity*.

One example cut from the 2015 (and last) ITRS report

DRAM products are approaching fundamental limitations as scaling DRAM capacitors is becoming very difficult in 2D structures. It is expected that these limits will be reached by 2024 and after this year DRAM technology will saturate at the 32Gbit level unless some major breakthrough will occur.
Flash memory on the other hand ...

Meanwhile, on the demand side

... the industry is busy creating the next growth bubble.



The line **4.6x / year** seems well below the current trend...

Source: epoch.ai

Evolution of Generative Pre-trained Transformers (GPT) in OpenAI

Model	Architecture	Parameter count	Training data	Release date	Training cost
GPT-1	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax.	117 million	BookCorpus: 4.5 GB of text, from 7000 unpublished books of various genres.	June 11, **2018**	"1 month on 8 GPUs", or 1.7e19 FLOP.
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, **2019** (initial/limited version) and November 5, 2019 (full version)	"tens of petaflop/s-day", or 1.5e21 FLOP.
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion	499 Billion tokens consisting of CommonCrawl (570 GB), WebText, English Wikipedia, and two books corpora (Books1 and Books2).	May 28, **2020**	3640 petaflop/s-day, or 3.2e23 FLOP.
GPT-3.5	Undisclosed	175 billion	Undisclosed	March 15, **2022**	Undisclosed
ChatGPT	Undisclosed	? (rumor 20M???)		**November 20, 2022**	
GPT-4	Also trained with both text prediction and RLHF; accepts both text and images as input. Further details are not public.	Undisclosed (1.8 trillon aka 1.8e12)	Undisclosed (13 trillon tokens, aka 1.3e13)	March 14, **2023**	Undisclosed. Estimated 2.1e25 FLOP.

From https://en.wikipedia.org/wiki/Generative_pre-trained_transformer

Compute requirement

~ x 88

~ x 213

~ x 65

~ x 1 218 360

13

1218360x in 5 years is about **16x / year**...

Source: Marc Duranton + wikipedia

Visualizing $1.8 \cdot 10^{12}$ parameters

- numbers printed in 100 lines of 5 columns on 2 sides of an A4 paper:
  1000 numbers/sheet
- one pack of 500 sheets is about 5 cm, so about 1000 sheets/10cm, or 10^{10} numbers/m
- $1.8 \cdot 10^{12}$ parameters is 180m of library shelves.

Not that bad for a summary after reading all the internet...

This was just about training...

Serving millions of users also consumes resources...

From `https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use`

- 1 typical GPT4 query is estimated to cost 10^{14} FLOP.
- which would consume 0.3Wh on modern GPUs.

(discounting the network cost of course)

This sounds very small

0.3Wh, or 1000 J, is the energy consumed by your laptop in a few seconds.

This was just about training...

Serving millions of users also consumes resources...

From `https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use`

- 1 typical GPT4 query is estimated to cost 10^{14} FLOP.
- which would consume 0.3Wh on modern GPUs.

(discounting the network cost of course)

This sounds very small

0.3Wh, or 1000 J, is the energy consumed by your laptop in a few seconds.
It is also the energy of a 1kg mass falling from 100 m above.

This was just about training...

Serving millions of users also consumes resources...
From `https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use`

- 1 typical GPT4 query is estimated to cost 10^{14} FLOP.
- which would consume 0.3Wh on modern GPUs.

(discounting the network cost of course)

This sounds very small

0.3Wh, or 1000 J, is the energy consumed by your laptop in a few seconds.
It is also the energy of a 1kg mass falling from 100 m above.

The problem is the exponential growth of the number of idiots.

- ChatGPT alone, in 2024, produces an estimated 200B words per day

Discrepancy with Moore's low is elegantly solved
by an exponential growth of datacenters and coal-fuelled energy.

Conclusion

Nothing to be pessimistic about

... Your jobs as computer engineers are not in danger (so focus on climate and biodiversity)

Nothing to be pessimistic about

... Your jobs as computer engineers are not in danger (so focus on climate and biodiversity)

- Current level of integration is here to stay. Maybe we won't get 2x better processors in two years, but we will still get current processors.
  (conversely, after peak oil, we get exponentially less petrol each day!
  This end-of-exponential will hurt much more.)

Nothing to be pessimistic about

... Your jobs as computer engineers are not in danger (so focus on climate and biodiversity)

- Current level of integration is here to stay. Maybe we won't get 2x better processors in two years, but we will still get current processors.
  (conversely, after peak oil, we get exponentially less petrol each day!
  This end-of-exponential will hurt much more.)
- Human I/O bandwidth is already saturated by the capacity of current tech.
  - digital sound quality improved drastically from square waves to stereo 16 bit @ 44KHz (reached in the 90s), and didn't progress since then.
  - digital video reached a similar plateau 10 years ago. Retina display etc.
  - we can still go real 3D, and then ?
  - (parallel: cars reached the max speed that humans can manage 80 years ago)

Nothing to be pessimistic about

... Your jobs as computer engineers are not in danger (so focus on climate and biodiversity)

- Current level of integration is here to stay. Maybe we won't get 2x better processors in two years, but we will still get current processors.
  (conversely, after peak oil, we get exponentially less petrol each day!
  This end-of-exponential will hurt much more.)
- Human I/O bandwidth is already saturated by the capacity of current tech.
  - digital sound quality improved drastically from square waves to stereo 16 bit @ 44KHz (reached in the 90s), and didn't progress since then.
  - digital video reached a similar plateau 10 years ago. Retina display etc.
  - we can still go real 3D, and then ?
  - (parallel: cars reached the max speed that humans can manage 80 years ago)
- There is space for negative growth.
  Fun game: count the MBytes transfered for 144 characters by Twitter,
  or for 10 lines of text in Marmiton.

Nothing to be pessimistic about

... Your jobs as computer engineers are not in danger (so focus on climate and biodiversity)

- Current level of integration is here to stay. Maybe we won't get 2x better processors in two years, but we will still get current processors.
  (conversely, after peak oil, we get exponentially less petrol each day!
  This end-of-exponential will hurt much more.)
- Human I/O bandwidth is already saturated by the capacity of current tech.
  - digital sound quality improved drastically from square waves to stereo 16 bit @ 44KHz (reached in the 90s), and didn't progress since then.
  - digital video reached a similar plateau 10 years ago. Retina display etc.
  - we can still go real 3D, and then ?
  - (parallel: cars reached the max speed that humans can manage 80 years ago)
- There is space for negative growth.
  Fun game: count the MBytes transfered for 144 characters by Twitter,
  or for 10 lines of text in Marmiton.
- There is space for better chips, better languages, better software... with the same transistors. This is what they do in the other departments of INSA, after all.

What disaster then ?

Disaster to those whose business model is based on Moore's law.

Shock, horror, PCs would become like fridges or dishwashers !

Old ones don't have more capacity or more performance than new ones.
We only buy a new one when the old one is broken.

- Some business models collapse (e.g. Microsoft doubling, from one Windows version to the next, the memory required to just do nothing)
  This is why they try to move to a subscription-based model.
- Some of the markets tries to move to data-oriented economy,
  with the admirable belief that data may grow exponentially forever.
- all this is not my speciality.

Pleasant times to be a computer engineer

- Admit it: all these years,
  software engineers have been parasites on the back of the electronic engineers.
  - *What Moore giveth, Gates taketh away*
- Now at last society needs you.

Pleasant times to be a computer engineer

- Admit it: all these years,
  software engineers have been parasites on the back of the electronic engineers.
  - *What Moore giveth, Gates taketh away*
- Now at last society needs you.

This conclusion is a failure, it only addressed societal impact on yourselves
(elite students in a rich country).

Try again.

Pleasant times to be a computer engineer

- Admit it: all these years,
  software engineers have been parasites on the back of the electronic engineers.
  - *What Moore giveth, Gates taketh away*
- Now at last society needs you.

This conclusion is a failure, it only addressed societal impact on yourselves
(elite students in a rich country).

Try again.

- Maybe we can use the end of Moore's law to transition from More to Better.

Pleasant times to be a computer engineer

- Admit it: all these years,
  software engineers have been parasites on the back of the electronic engineers.
  - *What Moore giveth, Gates taketh away*
- Now at last society needs you.

This conclusion is a failure, it only addressed societal impact on yourselves
(elite students in a rich country).

Try again.

- Maybe we can use the end of Moore's law to transition from More to Better.
- Maybe even beyond computer science.