Christof Teuscher

ECE 410/510: Hardware for AI and ML

Welcome and Introduction

Portland State University
Department of Electrical and Computer Engineering (ECE)
www.teuscher-lab.com
teuscher@pdx.edu





Cerebras Wafer-Scale Engine

Fabrication process	5nm
Silicon area	46,225mm²
Transistors	4 Trillion
AI-optimized cores	900,000
Memory (on-chip)	44GB
Memory bandwidth	21PB/s
Fabric bandwidth	214Pb/s

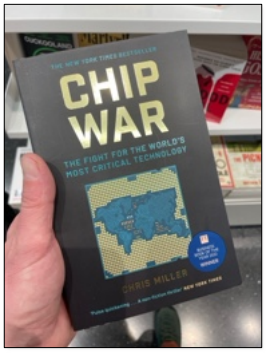
Cerebras WSE-3
4 Trillion Transistors
46,225 mm² Silicon

Electronic chips: the most complex human-made systems ever created

Largest GPU
80 Billion Transistors
814 mm² Silicon

cerebras

The human brain consists of 100 billion neurons and over 100 trillion synaptic connections.



Commodore C64

1982

Radio-frequency identification (RFID) tag

Same amount of memory!

The way we build, program, and use computers has fundamentally changed

MOS 6510, 1 MHz, 64 kB RAM, ~9000 transistors

Daily Mail

The perfect piece of toast: Scientists test 2,000 slices and find 216 seconds is the optimum time

By DAILY MAIL REPORTER
UPDATED: 04:03 EST, 22 July 2011

Share 136 View comments

Scientists today revealed the mathematical formula for a perfect slice of toast, showing that it is best cooked for exactly 216 seconds.

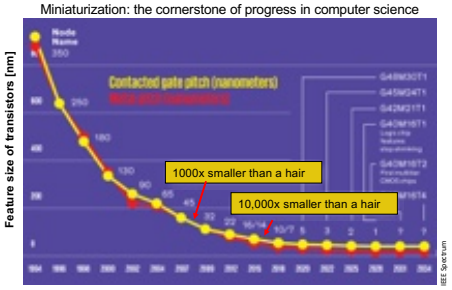
A team of researchers carried out a study which found the optimum thickness is 34mm and the ideal amount of butter is 0.44 grams per square inch.

The recommended cooking time gives the slice a 'golden-brown' colour and the 'ultimate balance of external crunch and internal softness'.

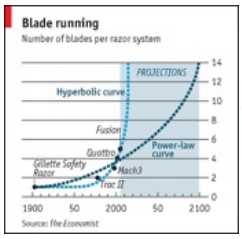
Perfect: The ideal slice of toast is cooked for just over three minutes, according to scientists

Smaller smaller smaller smaller smaller smaller smaller smaller

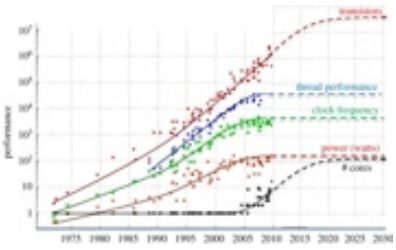
Miniaturization: the cornerstone of progress in computer science



The Cutting Edge: A Moore's Law for Razor Blades?



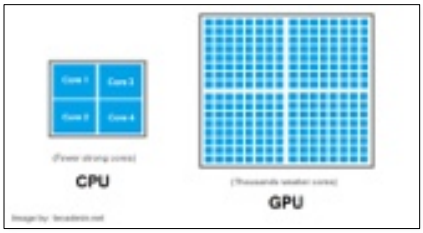
Source: The Economist, March 16, 2006

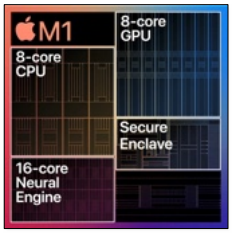


https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0061

https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0061

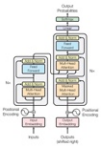
https://royalsocietypublishing.org/doi/10.1098/rsta.2019.0061





How did we get here?

Alan Turing

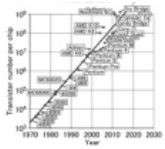


Deep neural networks

Transformers

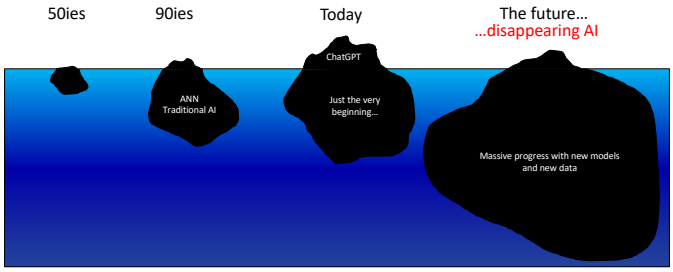
Models that learn and adapt from data

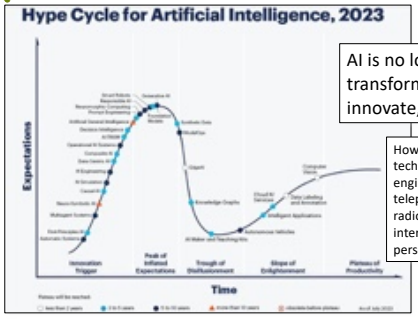
(Big) data



Moore's law

Where is this going?





AI is no longer hype—it's transforming how we work, innovate, and invest.

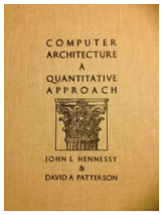
How does AI as a revolutionary technology compare to the steam engine, telegraph, photograph, telephone, automobile, airplane, radio, television, computer, the internet, mobile phone, and the personal computer?

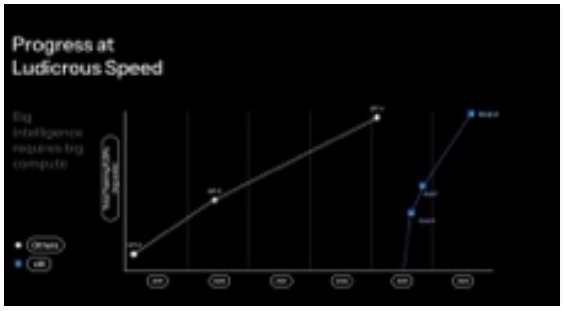
Traditional computer engineering and architecture is dead

AI/ML moves fast

- **"In AI, something breaks, ships, or goes viral every six hours."**
- Timescale: ~weeks

- For comparison: Hennessy & Patterson, first edition, 1990
- I have a beautiful 2nd edition from 1996 that I bought for my computer architecture class.





What challenges do computer engineers face?

- Deal with unseen complexities (server-scale AI machines, chiplets, trillions of transistors,…)
- Application-specific architectures for entirely new workloads
- Power barriers
- Shorter time to market
- Drastically increased need to keep up with state-of-the-art technological developments across the entire compute stack.
- Ethical issues
- …

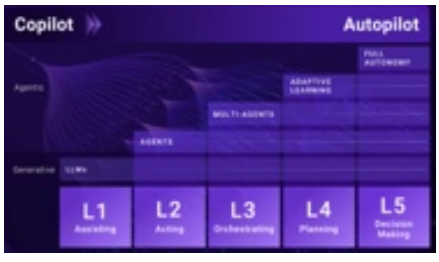
Synopsys lays out strategy for AI 'agents' to design computer chips

"[…] shift from designing single chips to AI server systems with hundreds or even thousands of chips in them while aiming to release a new server each year"

"These are very complex and difficult to design," Ghazi said of new AI computers. "The pressure engineers are feeling today is not only complexity, **it is complexity and the pace by when they need to deliver these products, as well as the cost**."

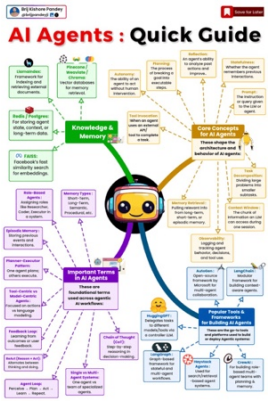
AgentEngineer: an "agent" that a human engineer can give instructions to.

https://www.reuters.com/technology/artificial-intelligence/synopsys-lays-out-strategy-ai-agents-design-computer-chips-2025-03-19/

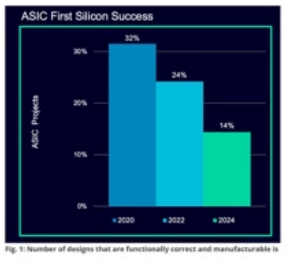


Evolution from generative to agentic AI. Source: Synopsys/SNUG

https://semiengineering.com/first-time-silicon-success-plummets



From LLMs to Agentic Systems



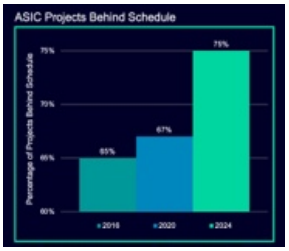


Fig. 1: Number of designs that are functionally correct and manufacturable is declining. Source: Siemens EDA/Wilson Research Group 2024 Functional Verification Study/DVCon

https://semiengineering.com/first-time-silicon-success-plummets

https://www.nytimes.com/2025/03/14/technology/why-im-feeling-the-agi.html

https://www.bbc.com/news/articles/c0kgydkr516o

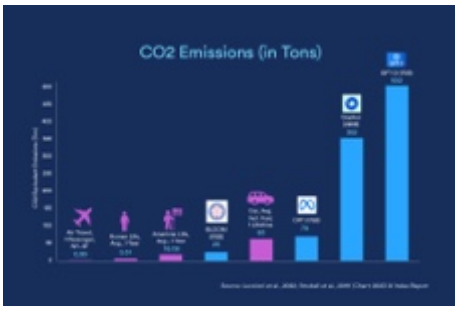
- LLMs prevent you from thinking and writing.
- You can't just learn how to craft a prompt for an A.I. chatbot without first having the experience, exposure and, yes, education to know what the heck you are doing.
- Learning is a messy, nonlinear human development process that resists efficiency. A.I. cannot replace it.

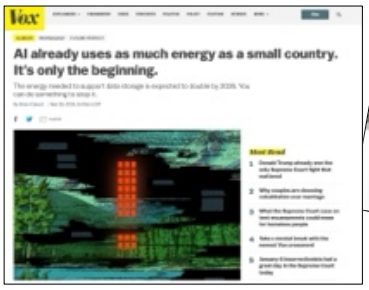
https://www.nytimes.com/2025/03/29/opinion/ai-tech-innovation.html

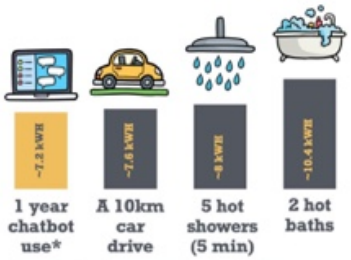
- Using LLMs leads to developing shallower knowledge compared to standard web search.
- "Processed" knowledge delivered to the learner, even if perfectly summarized, will often lead to shallower learning.
- Active learning is the best learning.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5104064

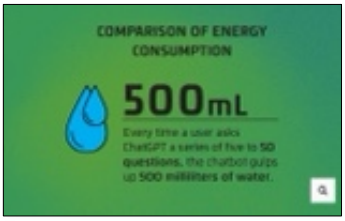
https://www.nytimes.com/2025/03/25/business/economy/white-collar-layoffs.html





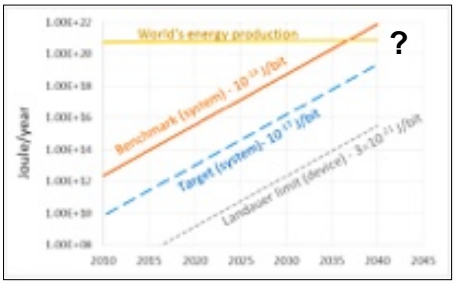


https://engineeringprompts.substack.com/p/ai-energy-use



https://wired.me/science/energy/ai-vs-bitcoin-mining-energy

https://www.oregonlive.com



Estimated total energy expenditure for computing, directly related to the number of raw bit transitions. Source: SIA/SRC.

Progress in AI/ML is currently largely dependent on advances in hardware

Thus this course.
Thus why this is a timely topic of interest to any computer engineer.

Christof Teuscher • teuscher@pdx.edu www.teuscher-lab.com/teaching

Co-design

https://coinflipscomputing.org