Foundations of Computing (at Scale)

...

Max Hawkins

Computing

Information

Definition (a single choice of many from Merriam-Webster):

> “knowledge obtained from investigation, study, or instruction”

Examples:

- Binary data in memory
- Output of an algorithm
- Collection of webpage visits
- Feeling of heat from the stove
- Sight of a flower

Information Processing is Critical

Examples with computing are bountiful and pervasive

- Usually dictated by Input -> Processing -> Output

Biological beings (like us) are doing this all the time too

- Feel hunger. Look in fridge for food. Open pickle jar. Eat pickle. ... Feel full.
- Each species' experience is affected by its umwelt
  - Definition: The specific way organisms of a particular species experience the world
  - Book recommendation: *An Immense World* by Ed Yong

Standardizing Common Information Processes

- Most humans have a general notion of less, more, equal, sets, etc.
  - A toddler recognizes when you steal all its cheerios
- However, settling a bill would be harder if both sides didn't share understanding

Introducing...MATH!

- Math serves as a common representation for many useful things
  - Common standards -> Easier communication and collaboration
  - Number systems, operations, written form (mostly standardized)
- Information Processing using Math: Computing
  - Definitions:
    - To determine especially by **mathematical** means
    - To determine or calculate by means of a **computer**

Computation

- Computation = Information Processing + Math?
- Humans evolved in a world without our explicit formulations of math
  - —> The human brain is pretty bad at math
  - Maximum working memory of 5 - 9 ‘numbers’ at any time
  - Forgetful, imprecise, hard to scale, unreliable, and long manufacturing/training time...
- Pen/paper, the abacus, and calculators are good but not great

- Information Processing is Valuable -> Design an Optimized Tool:

The Computer?

Computer Building Blocks

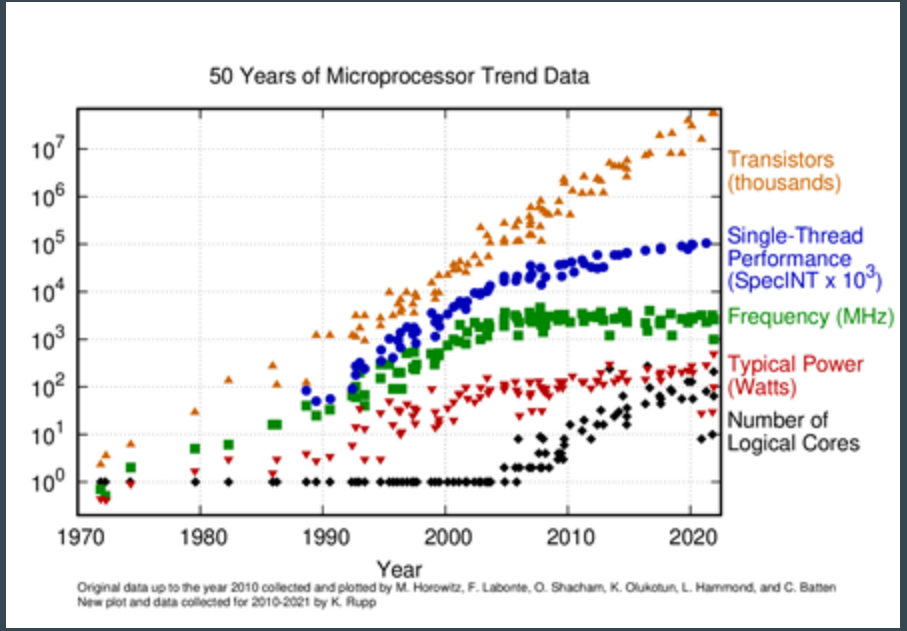
- Digital - Individual states of 0 or 1 (binary digits or bits)
- Electrical - Uses electricity/voltage levels for this encoding

- Vacuum tubes (too big and energy intensive)
- Transistors
  - Fundamental building block of digital computing for the past ~70 years

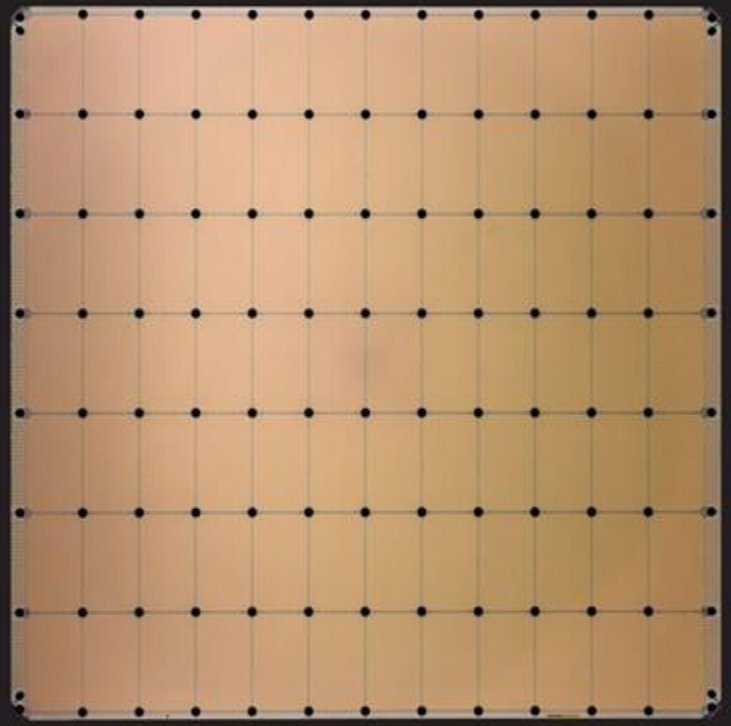
Computing at Scale

Scaling

- Transistors got smaller, cheaper, and more power efficient
  - → Computing performance increased
- Moore's law:
  - The number of transistors on a microchip doubles every ~2 years
  - How does transistor density play a role?
- Dennard Scaling:
  - Power density of transistors stays constant with size
  - Same unit chip area → Same power draw
- Where does this take us?



Cerebras Wafer-Scale Engine

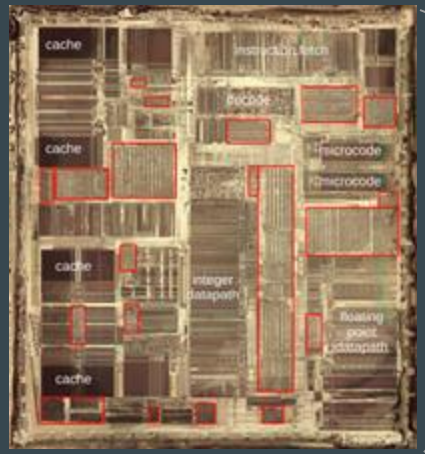


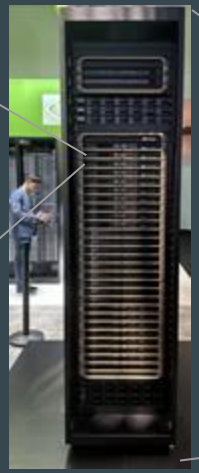
Cerebras WSE-2 7nm
2.6 Trillion Transistors
46,225 mm² Silicon

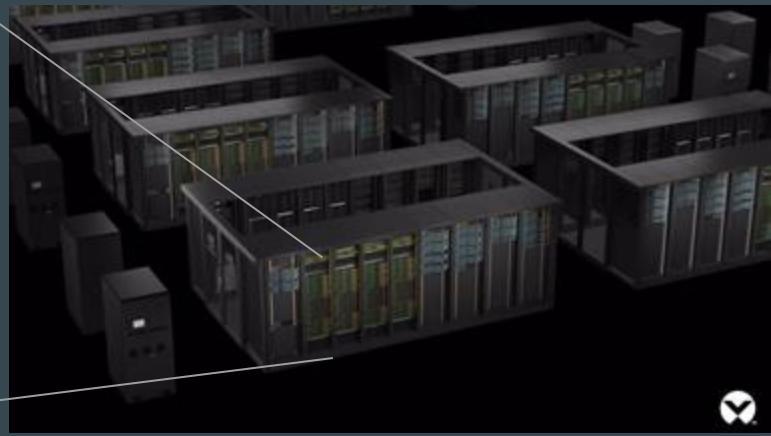
Largest GPU
54.2 Billion Transistors
826 mm² Silicon

Still Not Enough?

- Bigger, faster, more expensive, adds more cores
- Not just one device on a system -> local interconnect (PCIe)
- Dual socket motherboards
- Specialized hardware (GPUs, TPUs, DPUs, etc)
- Scale out → Multiple nodes → Interconnect
- Transistor, chiplet, chip, device, node, rack, aisle, datacenter, ...
    - From atomic to *global* scales







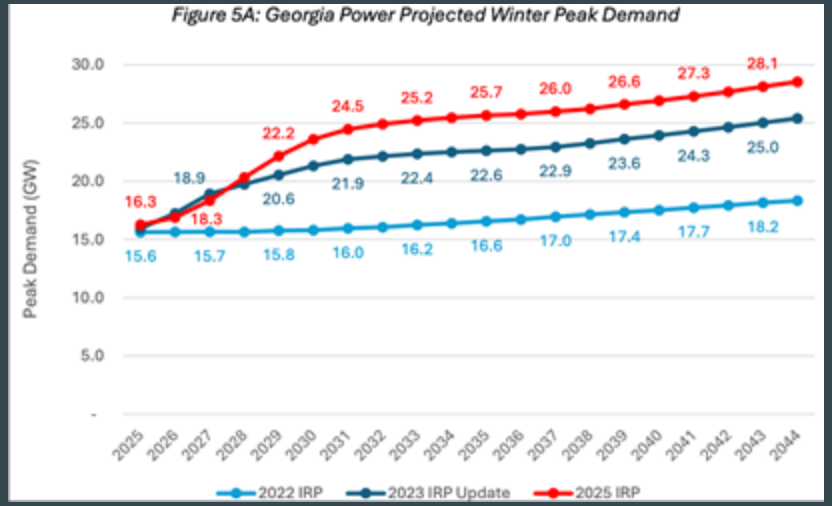
1. *Ken Shirriff's blog*
2. *TechPowerUp*
3. *The Register*
4. *Engineering News*
5. *Data Center Dynamics*

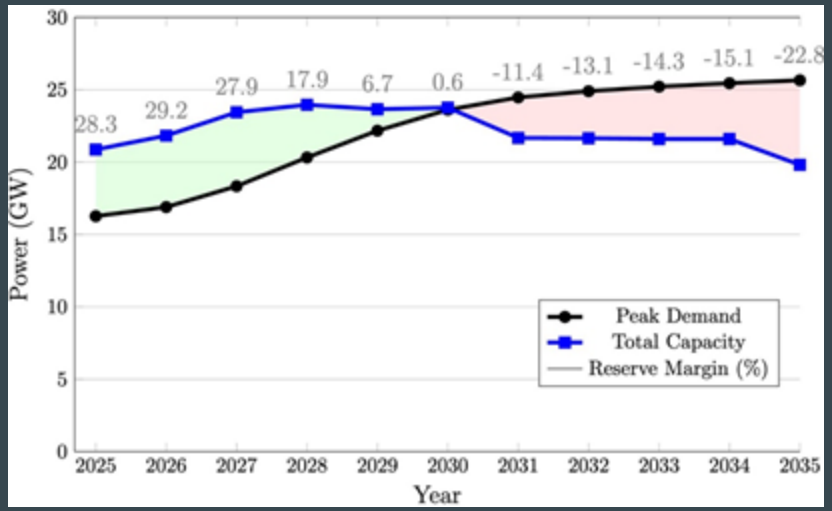
Unprecedented Scale and Impact

- Before AI, largest supercomputers ~20 MW
  - US Federal HPC systems like Frontier, Aurora, and El Capitan
- Meta plans to build a 5 GW AI DC (Hyperion)
  - 250 times larger!
- A traditional nuclear power plant outputs ~ 1 GW of power
- Massive Power + Inefficiencies = Large Cooling Demand
  - Evaporative water cooling is effective but water consuming
  - Example: 1 DC in Georgia will evaporate up to 6 million gallons of water per day
    - Equivalent to ~600-700 thousand average US people

Power Impacts

- AI demand is causing surging power demand
- Right: Georgia Power data
  - Top: The demand increases came quick
  - Bottom: Insufficient power generation
- Also, AI training power swings
  - At scale, training power can fluctuate 100s of MWs in seconds
  - Our grid wasn't built for this
- AI inference has a similar effect
  - Prefill vs decode phases
  - But the scale of inference synchronization is much smaller → negligible impact





Water Impacts

- Secondary water consumption: Energy generation
  - Thermoelectric power generation is the largest consumer of water in the US
  - Need to factor in water consumption of the power sources
- What happens in a drought?
- Who gets priority to water?
  - Example: That Georgia DC mentioned before is more than doubling its county's water consumption
- Are data centers 'critical' infrastructure like hospitals?
  - Less constraints imposed during water scarcity
- What are the tradeoffs of building a water-efficient data center?
  - Power, cost, efficiency, etc

Carbon Footprint

- Building, operating, and decommissioning DCs all have environmental impacts
- AI demand surge → Rush to build DCs → Build energy plants quick
  - Natural gas plants are in demand due to large power output and relatively quick build-out
- AI has a large carbon footprint

Future of Computing at Scale

- What is the limit to scaling computing?
  - Is there one?
  - Should there be?
- Jevons Paradox
  - Increasing resource efficiency induces increased demand and overall resource consumption
  - We have 70 years of data demonstrating this effect on computing
  - NOT like light production (e.g. candles to LEDs) where our eyes limit total demand
- What is your personal value weighting of computing?
  - How much are you willing to spend ($, power, water, carbon, time, etc) for AI/HPC work?

Computing is valuable,
operating at immense scale,
and its future can be shaped by you.

What future do you want to live in?