

"Computational Power and AI"

By Jai Vipra + Sarah Myers West AINOW.

Prof. David Andrews
Rm 527 JBHT

https://ainowinstitute.org/wp-content/uploads/2023/09/AI-Now_Computational-Power-an-AI.pdf

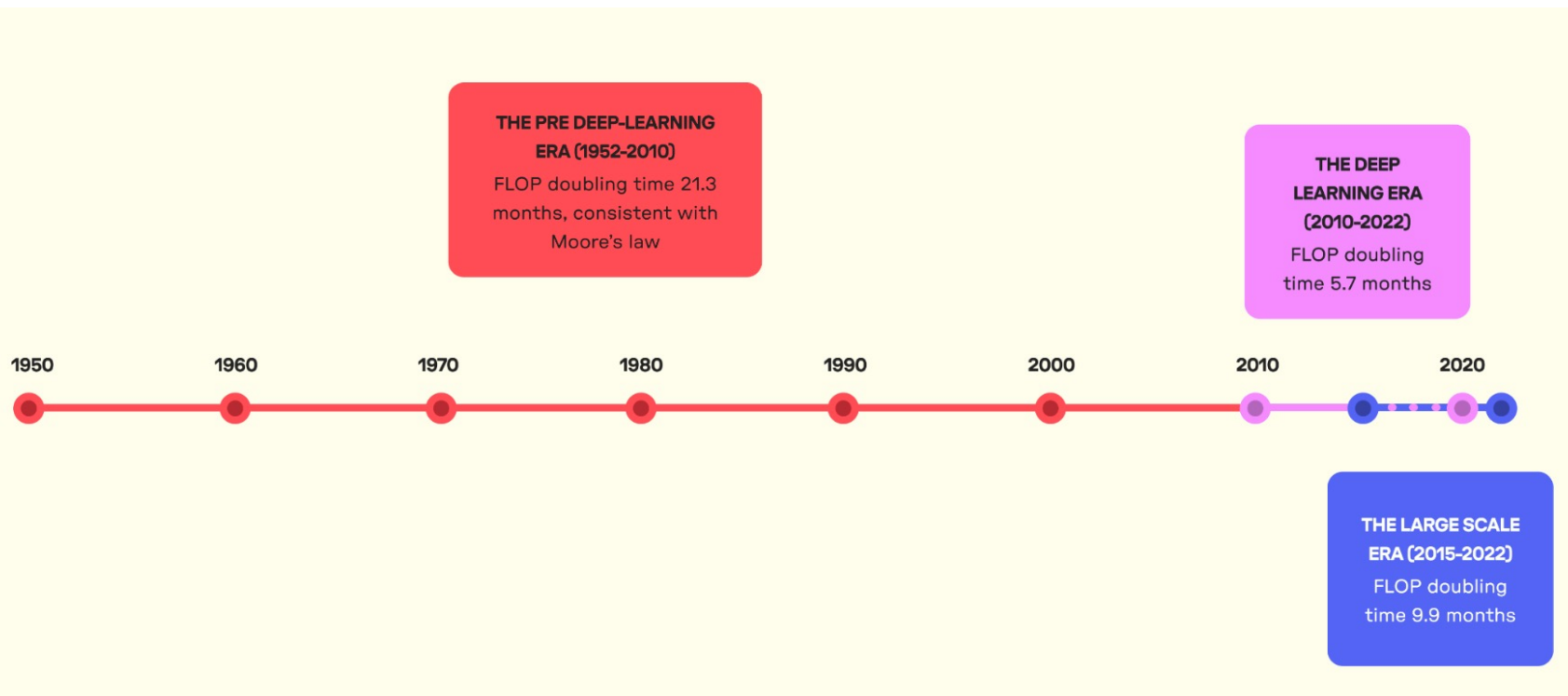


Approach to AI to date:

-Researchers in AI have largely concluded that increasing scale is key to increasing accuracy and performance in training deep learning models. This has driven an exponentially growing demand for compute power..



Compute Demand Growth



Pre Deep Learning Era: Compute for AI Models Doubled ~ 21 Months

Deep Learning Era (~2010): Doubled ~5.7 Months

Since 2015: Large-Scale Models Doubled ~9.9 Months

Regular-Scale Models Doubled ~5.7 Months



Defining Compute

1. **chips**, such as Graphics Processing Units (GPUs), which we will examine in detail in a later section;
2. **software** to enable the use of specialized chips like GPUs;
3. **domain-specific languages** that can be optimized for machine learning;
4. **data management software**; and
5. **infrastructure** in data centers that allows the use of thousands of chips together, including cabling, servers, and cooling equipment.



Core Thesis of the Article

- Compute is not just a technical input, it is a **structuring force** for:
 - Market power
 - Research direction
 - Environmental impact
 - Governance and policy
- Current approach to AI infrastructure self perpetuating, not sustainable long term.
- New Policies may be required to prevent hurtful monopolies



AI Dominance

- Compute is a predominant factor driving the industry today. Industry spends more than 80% of total capital spent on compute resources.
- Incentivizes cloud infrastructure providers to act in ways that protect their dominant position in the market
- Encourages lock-in into their cloud ecosystems.
- Reinforces the control of firms that already dominate the tech industry.

The rich get richer.....



Effects of Demands

- Demand for Limited Supply of Chips Extremely high
 - Orgs set up to provide GPU rental services
 - Purchases of GPUs by nation-states seeking competitive advantage
- Puts a few in Dominant Positions:
 - Cloud infrastructure firms (AWS, Google Cloud, MSAzure)
 - Chips: NVIDIA
 - Chip Fab: TSMC
- And unsustainable environmental issues
 - AI data center energy growth 160% by 2030. ~4.5% of global energy
 - TSMC accounts for 4.8% of Taiwan's national energy consumption, more than the entire capital city of Taipei



Self Perpetuating Cycles

- Accepted Scaling Assumption
- Bigger models + more compute => better results
- Industry consensus shaped investment
- Reinforces large-scale, centralized approaches



Hardware Lottery

- A research idea wins because it is the most suited to the available hardware and software.
 - Can stifle innovation and alternative approaches
 - Danger that this will reinforce an approach that is not long term sustainable in the long term

Cost of straying from mainstream hardware compatibility increases with time



Hardware-Software synergy

- Software Infrastructure Considerations
 - Corporate Allegiance to the status quo
 - Businesses invest and rely on stable software infrastructure
 - Resistance to retraining software developers
 - Resistance to "revving" up legacy code
 - Resistance to purchasing new Software Infrstr
- Software lock-in reinforces hardware dominance
 - GPU as ML ISA
 - CUDA defines algorithm design
 - Resists Growth of Alternative Algorithmic Approaches
 - Corporations resist switching to alternative hardware base.



Open Source Initiatives.....

- PyTorch opensource initiative under Linux foundation funded by Meta
- ONNX (Open Neural Network Exchange) opensource format for AI models to enable interoperability between frameworks (such as PyTorch and TensorFlow).



And the cycle continues

- More Demand causing shortage of GPUs
- Start-ups enter AI race by making contractual arrangements with Big Tech firms.
 - Grows dominance of existing firms
 - Google touted 70% of generative AI startups use Google cloud facilities.
- Big Tech Firms
 - Need for more compute shaping future product decisions.



Large-Scale AI Models

- How Much Compute and What Does it Cost
- At current growth rate: Compute costs in excess of entire US GDP by 2037
- Training GPT-4 "Probably" More than ~\$100 Million



Chips for large-scale AI

- GPUs
 - Dominance in Training
- Field Programmable Gate Arrays (FPGAs)
 - Arguments for:
 - Energy Efficiency
 - Flexibility
 - Arguments Against:
 - Lower Density/Performance
 - Lack of Software Centric Protocol Stack
- ASICs
 - Cost
 - Software Infrastructure
 - Time to Market



Components of Computer Hardware

- Logic
- Memory
- Interconnect

In traditional CPUs, memory tends to account for over half the cost of a server setup. (badly worded in paper). The increased costs of GPUs makes memory seem cheaper 😊



Supply Chain for AI Hardware

CHIP DESIGN	CHIP FABRICATION	DATA CENTERS
<ul style="list-style-type: none">• Nvidia• AMD• Intel• Arm• Broadcom	<ul style="list-style-type: none">• Taiwan Semiconductor Manufacturing Company• Samsung• Intel	<ul style="list-style-type: none">• Google• Amazon Web Services• Microsoft• Oracle• CoreWeave• Lambda Labs

Google: TPU used to develop Google Gemini: software/hardware integration impacts entire ecosystem leading to stronger monopolization

Microsoft: Athena => Maia 100 => Maia 200

3 nm node FP8/FP4 cores, 216GB HBM3e at 7 TB/s and 272MB of on-chip SRAM

AWS: Trainium



Market Dominance

- Data Centers:
 - NVIDIA 92 -98% of Data Center Market
 - AMD 8- 2%
 - Intel < 1%
- Overall GPU Market:
 - NVIDIA 94%
 - AMD ~6%



New Entries

- Cerebras: Wafer Scale Integration
 - Bandwidth 10,000x wrt NVIDIA GPUs
- Rain: Neuromorphic Architecture
- Groq; "Partnership" with NVIDIA
 - \$20 billion licensing deal to dominate real time AI market
 - Secure Language Processing Unit (LPU) architecture: low power/low latency user facing inferencing
 - Secure top talent
 - Hedges against HBM shortages (Groq uses SRAM)
 - Eliminates Groq as competitor



Investments

COMPANY	CLOUD COMPUTING MARKET SHARE (Q1 OF 2023)
Amazon Web Services	32 percent
Microsoft Azure	23 percent
Google Cloud	10 percent

Adapted from source¹²⁴

- Microsoft invested in OpenAI, Azure exclusive provider for OpenAI.
- Microsoft AI Supercomputers built for OpenAI



Investments

COMPANY	CLOUD COMPUTING MARKET SHARE (Q1 OF 2023)
Amazon Web Services	32 percent
Microsoft Azure	23 percent
Google Cloud	10 percent

Adapted from source¹²⁴

- Google Brain Team and DeepMind fully integrated into Google DeepMind.
- Amazon entered into partnerships with open source model developers and platforms.



Investments

COMPANY	CLOUD COMPUTING MARKET SHARE (Q1 OF 2023)
Amazon Web Services	32 percent
Microsoft Azure	23 percent
Google Cloud	10 percent

Adapted from source¹²⁴

- Oracle offers compute credit to AI startups
- Cloud companies invest in AI service companies



Data Center Growth

- Over 3,000 new data centers planned or under construction in U.S. circa 2025
- Examples:
 - Amazon \$11B Indiana
 - \$1.3B campus in Austin (2 GW)
 - Microsoft 1.5 GW facility in Wisconsin
 - OpenAI and Oracle Michigan
 -and others.....
- Worldwide
 - Microsoft \$3.5B Germany
 - Google \$1B in new UK facility
 - Latin America
 - Saudi Arabia



How to Reduce Compute Costs

- Don't Count on Moore's Law 😊
- New Algorithms: Smaller Models
- Paradigm Shifts and Breakthroughs
 - Memory bottleneck: New Memory Technology could provide significant breakthrough.
 - New Heterogeneous Technologies: Nanotech + CMOS
 - New Architectures: Processor in Memory (PIM), Neuromorphic, Quantum, analog computing architectures



Policy considerations

- Antitrust: Continued Vertical Integration reinforces Hardware Lottery grip:
 - Separate Cloud Provision from Chip Design
 - Separate Hardware from Software, or mandate interoperability
 - Separate AI Model Development from Cloud Infrastructure
 - Institute Nondiscrimination or Common Carrier Obligations Across Tech Stack
 - Prevent Further Market Concentration
 - Investigate Anticompetitive Conduct
 - Apply Existing Antitrust Principles to AI Compute Markets



Data Minimization

- Data and Compute are separate inputs to AI.
 - Scaling laws limit amount of data that can be efficiently used with a given amount of compute, exclusive data access becomes increasingly important as freely available internet data runs out.
 - Freely available data already used
 - Newly produced data is starting to be protected by more platforms.
 - Reddit, Twitter have implemented protections against free use of data from their platforms.
 - Relative value of internet data declining as it features more AI-generated content.



Data Minimization

- Embrace data-minimization mandates, prohibit collection/processing of sensitive data
- Prohibit secondary use of data collected from consumers for training AI models as violation over consumer control of personal data.



Discussion Questions

- Should compute be regulated as infrastructure?
- How do we promote/fund architectural diversity?

