AI, Solarpunk, and an Uncertain Future in Computing

rolltime (*she/her*)

July 12, 2024

Licensed under CC-BY-SA

2024-07-12

Al, Solarpunk, and an Uncertain Future in Computing

AI, Solarpunk, and an Uncertain Future in Computing

rolltime (ahe/her) July 12, 2024 Licensed under CC-BY-SA

Thank people, now that they're placated, let's talk about me!

Introduction

2024-07-12

Al, Solarpunk, and an Uncertain Future in Computing $\hfill \hfill \Box$ Introduction

Introduction

- I'm rolltime!
- Hacker, Software Engineer, Optimist

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing $\hfill \hfill \hfi$

I'm rolltime!
 Hacker, Software Engineer, Optimist

Recently finished my BA, wrote my thesis on Solarpunk and AI, I'm the lead engineer at a startup doing neither of those things.

Now I want to bring my findings here. It's my hope (ha) that you all leave this talk with newfound understanding and passion. This is my second HOPE talk.

Activity Pub

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing $\hfill \hfill \hfi$

⊲⊳ ActivityPub

I spoke about ActivityPub, the federated social networking protocol which was thrust into the spotlight following Elon Musk's acquisition of Twitter. I spoke about my vision for a better kind of social media and how I felt we could get there. The talk was fairly popular and got a lot of feedback. I really value feedback so I'd like to take the time to review some of that feedback now.

() @re

@rednafi 1 year ago Thanks for explaining the core architecture of a federated system in a single slide. This was incredible helpful.

台 11 🖓 🛛 Reply

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing $\hfill \hfill \hfi$

 $\bigoplus_{d\,=\,0^{-1}}^{detter the matrix and the second second$

Ørednafi: Thanks for explaining the core architecture of a federated system in a single siide. This was incredible helpful.



@forrest_wilkins 1 year ago Loved this, great talk!

占 🦓 Reply

AI, Solarpunk, and an Uncertain Future in Computing $\hfill \hfill \hfi$

@foreat_wilkins 1 year age Loved this, great talk!

@forrest_wilkins: Loved this, great talk!

S @SegFault01 11 months ago (edited)
It's unfortunate that the speaker sprayed his political/ideological viewa on the audience but apart from that it was a good talk.
C 4 57 Reply

AI, Solarpunk, and an Uncertain Future in Computing

Segfault01: It's unfortunate that the speaker sprayed his political/ideological viewa on the audience but apart from that it was a good talk.

3 disebati l'incent franklade beben gi (3) nev v Janin

RADICAL SOCIAL MOVEMENTS

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing *RADICAL SOCIAL MOVEMENTS*

RADICAL SOCIAL MOVEMENTS

Radical social movements, ending capitalism, and curbing climate change. Segfault01, this one is on you.

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing *RADICAL SOCIAL MOVEMENTS*

@SegFault01

@SegFault01

• Our relationship with technology is broken

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -RADICAL SOCIAL MOVEMENTS

Overview

· Our relationship with technology is broken

-Overview

Not just technology as it exists, but how we develop it

- Our relationship with technology is broken
- A Key Example

Al, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -RADICAL SOCIAL MOVEMENTS

-Overview

Our relationship with technology is broken

Overview

7

- Our relationship with technology is broken
- A Key Example
- A Beautiful Future

AI, Solarpunk, and an Uncertain Future in Computing *RADICAL SOCIAL MOVEMENTS*

-Overview

2024-07-12

Our relationship with technology is broken
 A Key Example
 A Beautiful Future

Overview

- Our relationship with technology is broken
- A Key Example
- A Beautiful Future
- Real-World Applications

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -RADICAL SOCIAL MOVEMENTS

-Overview

 Our relationship with technology is broken A Beautiful Future

Overview

- Our relationship with technology is broken
- A Key Example
- A Beautiful Future
- Real-World Applications
- Challenging the AI Mainstream

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -RADICAL SOCIAL MOVEMENTS

-Overview

· Our relationship with technology is broken A Beautiful Future · Challenging the Al Mainstream

Overview

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing \square AI: The Neverending Bubble

AI: The Neverending Bubble

AI: The Neverending Bubble

A Very Brief History

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 └─AI: The Neverending Bubble

A Very Brief History

• 2016: RNNs, CNNs, GANs

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 AI: The Neverending Bubble

A Very Brief History

A Very Brief History

2016: RNNs. CNNs. GANs

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 └─AI: The Neverending Bubble

A Very Brief History

A Very Brief History

· 2016: RNNs. CNNs. GANs • 2017: Transformer Takeover

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAl goes for-profit

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 AI: The Neverending Bubble

└─A Very Brief History

A Very Brief History

· 2016: RNNs. CNNs. GANs · 2020: GPT-3 released. OpenAl goes for-profit

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAl goes for-profit
- 2021: AlphaFold 2

AI, Solarpunk, and an Uncertain Future in Computing \square AI: The Neverending Bubble

└─A Very Brief History

A Very Brief History

2016: RNNs, CNNs, GANs
 2017: Transformer Taksover
 2020: GPT-3 released, OpenAl goes for-profit
 2021: AlphaFold 2

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAl goes for-profit
- 2021: AlphaFold 2
- 2022: Stable Diffusion

AI, Solarpunk, and an Uncertain Future in Computing \square AI: The Neverending Bubble

└─A Very Brief History

2016: RNMs, CNMs, GANs
 2017: Transformer Takeover
 2020: GPT-3 released, OpenAl goes for-profit
 2021: AlphaFold 2

A Very Brief History

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2
- 2022: Stable Diffusion
- 2023: GPT-4, Claude 2, LLaMA

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -AI: The Neverending Bubble

└─A Very Brief History

· 2016: RNNs. CNNs. GANs · 2017: Transformer Takeove 2020: GPT-3 released. OpenAl goes for-profil 2021: AlphaFold 2

A Very Brief History

GPT-4 was the first robot to pass the bar exam, unless you count Alan Dershowitz.

- 2016: RNNs, CNNs, GANs
- 2017: Transformer Takeover
- 2020: GPT-3 released, OpenAI goes for-profit
- 2021: AlphaFold 2
- 2022: Stable Diffusion
- 2023: GPT-4, Claude 2, LLaMA
- 2024: GPT-4o, Claude 3.5, AlphaFold 3

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -AI: The Neverending Bubble

└─A Very Brief History

A Very Brief History

· 2016: RNNs. CNNs. GANs · 2017: Transformer Takeover 2020: GPT-3 released. OpenAl goes for-profit · 2021: AlphaFold 2

Where We Are Now

• Money is still raining from the sky

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing \square AI: The Neverending Bubble

Where We Are Now

Al Ventures Attracted 96 Billion Dollars of Investment in 2023. Crypto: 30 Billion in 2021. Al: 42 Billion in 2021.

Where We Are Now

Where We Are Now

- Money is still raining from the sky
- The shovel-selling business is good

Al, Solarpunk, and an Uncertain Future in Computing $\hfill \Box$ Al: The Neverending Bubble

└─Where We Are Now

nvidia briefly became the world's most valuable company

Where We Are Now

Money is still raining from the sky
 The shovel-selling business is good

Where We Are Now

- Money is still raining from the sky
- The shovel-selling business is good
- Around 30% of people use ChatGPT weekly

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -AI: The Neverending Bubble

-Where We Are Now

ChatGPT had the second fastest user growth, behind threads. This number is probably higher in the US.

Where We Are Now

· The showel-selling business is good · Around 30% of people use ChatGPT week!

Staggering Growth



Al, Solarpunk, and an Uncertain Future in Computing \square Al: The Neverending Bubble

└─Staggering Growth



Staggering Growth



Al, Solarpunk, and an Uncertain Future in Computing $\hfill \Box$ Al: The Neverending Bubble

└─Staggering Growth



A Dawning Realization

• Al skepticism is becoming mainstream

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing —AI: The Neverending Bubble

└─A Dawning Realization

More people are now saying what artists have been saying for months. I will fucking piledrive you if you mention AI again - Ludicity

A Dawning Realization

I'm going to ask ChatGPT how to prepare a garotte and then I am going to strangle you with it, and you will simply have to pray that I roll the 10% chance that it freaks out and tells me that a garotte should consist entirely of paper mache and malice.

A Dawning Realization

- Al skepticism is becoming mainstream
- Even Sequoia says it's unsustainable!

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -AI: The Neverending Bubble

└─A Dawning Realization

Al will need to generate an additional \$600bn of revenue to become profitable

A Dawning Realization

· Even Sequoia says it's unsustainab

• Bad.

AI, Solarpunk, and an Uncertain Future in Computing $\hfill \Box$ AI: The Neverending Bubble

Energy Demand: Just How Bad Is It?

Billions of dollars into waste heat. One person flying from New York to LA results in a staggering two metric tons of CO2 emissions. Training GPT-4 cost OpenAI \$100m in hardware and electricity and produced the equivalent of 2500 of these flights, or 156 person-years of CO2 for the average American.

Energy Demand: Just How Bad Is It?

• Bad.

• Training GPT-4: 2500 Flights, 100 cars

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -AI: The Neverending Bubble

-Energy Demand: Just How Bad Is It?

Training is energy-intense, but depending on model lifecycle, inference can be an even bigger problem. ChatGPT consumed an estimated 564 MWh per day in early 2023 to serve 195 million requests. Accounting for user growth and larger models, a conservative estimate has ChatGPT producing 100 flights per day of CO2.

Energy Demand: Just How Bad Is It?

. Training GPT-4: 2500 Flights, 100 cars

Energy Demand: Just How Bad Is It?

2024-07-12

Al, Solarpunk, and an Uncertain Future in Computing \square Al: The Neverending Bubble

Energy Demand: Just How Bad Is It?

Bad.
 Training GPT-4: 2500 Flights
 ChatGPT: >100 Flights Per Day

• Bad.

- Training GPT-4: 2500 Flights
- ChatGPT: >100 Flights Per Day

• Bad.

- Training GPT-4: 2500 Flights
- ChatGPT: >100 Flights Per Day
- Google's carbon emissions up 48%

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -AI: The Neverending Bubble

-Energy Demand: Just How Bad Is It?

ChatGPT: >100 Elights Per D;

Energy Demand: Just How Bad Is It?

Google's carbon emissions are up 48% since 2019, mostly due to new datacenter demand. They have directly stated that AI is at least partially responsible; "reducing emissions may be challenging due to increasing energy demands from the greater intensity of AI compute".

They've also claimed that their Al-powered Search is ten times more expensive than the standard version. Based on that higher cost, we can estimate that if Google were to use AI for every search, the system's power usage might approach 30 TWh per year.

• Bad.

- Training GPT-4: 2500 Flights
- ChatGPT: >100 Flights Per Day
- Google's carbon emissions up 48%
- Al-Powered Google Search: 30 TWh

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -AI: The Neverending Bubble

-Energy Demand: Just How Bad Is It?

 Training GPT-4: 2500 Flights · Google's carbon emissions up 48% Al-Powered Google Search: 30 TWI

Energy Demand: Just How Bad Is It?

I've often heard it claimed that AI is using "as much power as a small country". That's actually false: it's as much as a large one. 30 TWh per year has Google alone tied with Bulgaria as the 66th largest consumer of electricity in the world, producing 1.8 megatons of co2 per year, or almost 2500 flights per day.

These numbers show a real, meaningful impact on the climate. That's scary, especially at a time when reducing carbon emissions should be humanity's number one priority. That raises the question: what do we do about it?

Moving Forward

• Is radical degrowth/deurbanization the solution?

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 AI: The Neverending Bubble

└─Moving Forward

Why not just ban AI?


Neo-Pastoral Anarcho-Primitivist Cottagecore Bullshit

Neo-Pastoral Anarcho-Primitivist Cottagecore Bullshit

- Unrealistic; if we all move to the country it won't be the country anymore.
- Density isn't inherently bad for the environment
- Long-time human urge for a "simpler, greener" time before cities. Cot-tagecore
- Impossible to disentangle that urge from racism. White flight. Debt spiral.
- Not the first time; look at Andrew Jackson.
- We need a different dream. We need to believe in a world where the practical merges with the beautiful. Where dense urban environments thrive with lush greenery and sustainable food systems. Where optimism isn't a fallacy, but the path to a brighter future. We need Solarpunk.

Solarpunk



Munashichi, Future Economic View of Innocence, 2015

AI, Solarpunk, and an Uncertain Future in Computing Neo-Pastoral Anarcho-Primitivist Cottagecore Bullshit Solarpunk



Solarpunk is what happens when we dare to ask the question: What if we could do better?

Solarpunk: An Introduction

• Response to Cyberpunk

AI, Solarpunk, and an Uncertain Future in Computing Neo-Pastoral Anarcho-Primitivist Cottagecore Bullshit Solarpunk: An Introduction

Cyberpunk is dystopian, solarpunk is hopeful

22

Solarpunk: An Introduction

- Response to Cyberpunk
- Beautiful solutions are better

Science fiction is a form of activism. Out of the box thinking

Solarpunk: An Introduction

Solarpunk: An Introduction

- Response to Cyberpunk
- Beautiful solutions are better
- Post-capitalist

Solarpunk: An Introduction AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 Neo-Pastoral Anarcho-Primitivist Cottagecore Bullshit Solarpunk: An Introduction

Solarpunk is counterculture; to progress is to move past competition and towards collaboration.

Response to Cyberpunk

Post-capitalist

What do you get when you combine radical optimism, out of the box thinking, and counterculture? What do you get when you form communities centered around bold new ideas and inspired problem solving? You get hackers! Solarpunk is how we hack our way towards a better world. So let's do it!

Solarpunks Make Beautiful Things



AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -Neo-Pastoral Anarcho-Primitivist Cottagecore Bullshit

-Solarpunks Make Beautiful Things



Stela Xhiku, Heliostat #1, 2023

Solarpunks Make Beautiful Things



AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -Neo-Pastoral Anarcho-Primitivist Cottagecore Bullshit

-Solarpunks Make Beautiful Things



Me, Farming Web of Trust, 2022

Solarpunks Make Beautiful Things



Sofi, Acorn Land Labs, 2023

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -Neo-Pastoral Anarcho-Primitivist Cottagecore Bullshit

-Solarpunks Make Beautiful Things



Solarpunks Make Beautiful Things

Let's look at a cool technological application of Solarpunk ideas.

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing \Box Let's hack the web! (the other way)

Let's hack the web! (the other way)

Let's hack the web! (the other way)

The Solar Protocol

Al, Solarpunk, and an Uncertain Future in Computing 2024-07-12 Let's hack the web! (the other way)

L The Solar Protocol

• Energy-Centered

The Solar Protocol

Al, Solarpunk, and an Uncertain Future in Computing \Box Let's hack the web! (the other way)

└─The Solar Protocol

The Solar Protocol

Energy-Centered
Planet-Scale

- Energy-Centered
- Planet-Scale

The Solar Protocol

- Energy-Centered
- Planet-Scale
- Naturally Intelligent

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing \Box Let's hack the web! (the other way)

└─The Solar Protocol

permacomputing, energy-centered design

The Solar Protocol

Energy-Centered
Planet-Scale
Naturally Intelligent



72hr power diagram for solarprotocol.net. Served from Trent University in Canada. How can we apply energy-centered design to AI?

AI, Solarpunk, and an Uncertain Future in Computing Let's hack the web! (the other way)

The prime degree for a state state from the bismuth

Servers in Canada, the US, China, Chile, Kenya, and more.

2024

Approximate Multipliers

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing $\hfill \square$ Approximate Multipliers

Summary of my BA thesis

Approximate Multipliers

"Efficient" Deep Learning: A Solarpunk Approach

• What do we mean by efficient?

AI, Solarpunk, and an Uncertain Future in Computing \square Approximate Multipliers

└─" Efficient" Deep Learning: A Solarpunk Approach

Reducing memory- reduces hardware cost. Reducing compute (FLOPs)increases speed and reduces power requirements. As Solarpunks, we're concerned with power. Let's look at some techniques from the AI mainstream.

"Efficient" Deep Learning: A Solarpunk Approach

· What do we mean by efficient?

Some Other Ideas

	Quant.	Pruning
Reduce memory	\checkmark	
Reduce FLOPs		\checkmark
Reduce power	\checkmark	\checkmark
Used in training		
Zero-shot		\checkmark
Equal Performance	\checkmark	\checkmark

AI, Solarpunk, and an Uncertain Future in Computing — Approximate Multipliers

└─Some Other Ideas

(Outet	Develop
Reduce memory	Z Z	- raning
Reduce ELOPs		7
Reduce power	7	-
Used in training		
Zero-shot		4
Equal Performance	4	4

Some Other Ideas

Quantization: Reduces bit width. Saves a lot of memory but for large models inability to represent small gradients is a problem. Often used after training. Pruning: Doesn't affect memory, simply sets some weights to zero to save compute. Sometimes retrained. What all of these have in common is they attempt to reduce the amount of multiplication done.

Let's Talk about Multiplication

• Consider long multiplication:

4	7	0	7	8	1	
4	3	4	4			
	3	2	5	8		
		3	8	0	1	
		\times	8	6	7	
			5	4	3	

AI, Solarpunk, and an Uncertain Future in Computing — Approximate Multipliers

Let's Talk about Multiplication

	et's	Talk	about	Multipli	ication
--	------	------	-------	----------	---------

Consider long multiplication:

5 4 3 × 8 6 7 3 8 0 1 3 2 5 8 4 3 4 4

Let's Talk about Multiplication

• Consider long multiplication:

			5	4	3	
		×	8	6	7	
		3	8	0	1	
	3	2	5	8		
4	3	4	4			
4	7	0	7	8	1	

• 99% of AI compute is multiplication!

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -Approximate Multipliers

-Let's Talk about Multiplication

alk about Multiplication
3801
99% of AI compute is multiplication!

Let's

repeated addition; quadratic complexity. Quantization and pruning reduce the amount of multiplication done. What if we could use hardware that made multiplication more efficient?

Log Multiplication

• Recall from HS math:

 $\log(a \cdot b) = \log(a) + \log(b)$

• Binary log is extremely fast and efficient, but has high error

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -Approximate Multipliers

-Log Multiplication

Log Multiplication

Logarithm product rule. This idea gives rise to Approximate Multipliers. AI workloads have been shown to be tolerant to small errors in multiplication. If we replace regular multiplication with log multiplication, we can reduce the power significantly.

Some Other Ideas

	Quant.	Pruning	AMs
Reduce memory	\checkmark		
Reduce FLOPs		\checkmark	
Reduce power	\checkmark	\checkmark	\checkmark
Used in training			\checkmark
Zero-shot	*	\checkmark	\checkmark
Equal Performance	*	\checkmark	\checkmark

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -Approximate Multipliers

-Some Other Ideas

	Quant.	Pruning	AMS
Reduce memory	4		
Reduce FLOPs		4	
Reduce power	4	4	4
Used in training			4
Zero-shot	•	4	4
Equal Performance	·	4	4

Some Other Ideas

AMS affect the underlying framework. They reduce the computational cost of multiplication. They can be used in concert with all the other techniques here. Equal or better performance. Energy-centered design

• Drop-in

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing — Approximate Multipliers

Using AMs in DNNs

No change in architecture, no retraining90+% energy savings.

Using AMs in DNNs

• Drop-in

• Minimal Performance Degradation

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 -Approximate Multipliers

└─Using AMs in DNNs

Stochastic regularization90+% energy savings.

Using AMs in DNNs

Minimal Performance Degradation

• Drop-in

- Minimal Performance Degradation
- Modular

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing \square Approximate Multipliers

└─Using AMs in DNNs

Usable alongside everything else we talked about 90+% energy savings.

Using AMs in DNNs

· Minimal Performance Degradation

• Drop-in

- Minimal Performance Degradation
- Modular
- Massive Energy Savings

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing \square Approximate Multipliers

└─Using AMs in DNNs

90+% energy savings.

Using AMs in DNNs

Drop-in
Minimal Performance Degradation
Modular
Massive Energy Savings

Case Study: AlexNet

AlexNet Image: 224 (height) × 224 (width) × 3 (channels) Convolution with 11 × 11 kernel+4 stride: 54 × 54 × 96 ReLu Pool with 3×3 max, kernel+2 stride: 26×26×96 Convolution with 5×5 kernel+2 pad:26×26×256 ReLu Pool with 3×3 max.kernel+2stride:12×12×256 Convolution with 3×3 kernel+1 pad:12×12×384 ReLu Convolution with 3×3 kernel+1 pad:12×12×384 ReLu Convolution with 3×3 kernel+1 pad:12×12×256 ReLu Pool with 3×3 max.kernel+2stride:5×5×256 flatten Dense: 4096 fully connected neurons ReLu, dropout p=0.5 Dense: 4096 fully connected neurons ReLu. dropout p=0.5 Dense: 1000 fully connected neurons Output: 1 of 1000 classes

AI, Solarpunk, and an Uncertain Future in Computing — Approximate Multipliers

Case Study: AlexNet



This is AlexNet, a well-known CNN which is complex enough to be a useful example but small enough to test easily. I implemented both precise and approximate versions and trained them on the CIFAR-10 dataset for 50 epochs.

-07-12

2024-

Results



Precise Train Accuracy Precise Validation Accuracy Approximate Train Accuracy
Approximate Validation Accuracy



AI, Solarpunk, and an Uncertain Future in Computing — Approximate Multipliers

└─ Results

07-12

2024



Results

Precise: Final validation 81.8, training 98.5 Approx: Final validation 81.7, training 93.4 Rates of learning are similar, AMs get off to a slower start.

• Nvidia's Monopoly

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing \square Approximate Multipliers

What's the Catch?

What's the Catch?

Nvidia's Monopo

- Nvidia's Monopoly
- Model Size/Speed

2024-07-12

AI, Solarpunk, and an Uncertain Future in Computing \square Approximate Multipliers

What's the Catch?

What's the Catch?

Nvidia's Monopoly
Model Size/Speed

- Nvidia's Monopoly
- Model Size/Speed
- Market Share

2024-07-12

Al, Solarpunk, and an Uncertain Future in Computing $\hfill \Box$ Approximate Multipliers

What's the Catch?

What's the Catch?

Nvidia's Monopoly
Model Size/Speed
Market Share

- Nvidia's Monopoly
- Model Size/Speed
- Market Share
- Consequences shouldnt be an externality!

Al, Solarpunk, and an Uncertain Future in Computing $\hfill \Box$ Approximate Multipliers

What's the Catch?

What's the Catch?

Nvidia's Monopoly
Model Size/Speed
Market Share
Consequences shouldnt be an externality!

Conclusion

2024-07-1

AI, Solarpunk, and an Uncertain Future in Computing —Conclusion

Why are we here? Because of the name of the conference: Hope. Solarpunk isn't about saving the world, but it is about trying our hardest to.

If the rest of this means nothing to you, look at me now and remember this: you are the future. Everyone in the room with me, everyone watching live, everyone viewing the recording; you are the future. The freaks and the hackers and the weirdos. We all have a role to play in saving the world. The more people believe that it's possible to change the world, the more likely it is to happen. So get out there and make it happen.

Conclusion

Thank You

2024-07-12

Al, Solarpunk, and an Uncertain Future in Computing $\hfill \Box$ Thank You

Thank You

Links and Such

@rolltime@freeradical.zone



rollti.me/hope2024

AI, Solarpunk, and an Uncertain Future in Computing Laboratory Thank You

inks and Such



Mitchell's Multiplier Example

Let's take the binary log of 18. 18 = 00010010
Find the most significant one bit, here 2⁴.
Take the position (4) as the significand, and the rest as the mantissa; b0.0010 = 0.125. Thus:

 $\log_2 18 \approx 4.125$

In fact it's about 4.17, so we got pretty close!

AI, Solarpunk, and an Uncertain Future in Computing

└─Mitchell's Multiplier Example

Computers- shift to left, counter, etc.

 $\log_2 18 \approx 4.12$

In fact it's about 4.17, so we got pretty closel

Mitchell's Multiplier Example

In Formal Terms

 Consider the binary representation B of an N-bit unsigned integer b_{N-1}b_{N-2}...b₁b₀:

$$B = \sum_{i=0}^{N-1} 2^{i} b_{i}$$
 (1)

Say the most significant one bit in B is at position k. B can then be written as:

$$B = 2^{k} \left(1 + \sum_{i=0}^{k-1} 2^{i-k} b_{i} \right)$$
(2)

Al, Solarpunk, and an Uncertain Future in Computing $\hfill \Box$ Thank You

└─In Formal Terms



In Formal Terms, Cont.

• Let x be:

$$x = \sum_{i=0}^{k-1} 2^{i-k} b_i \tag{3}$$

where $0 \le x < 1$. We can write:

$$B = 2^k (1+x) \tag{4}$$

By log rules, the accurate binary log of B is:

$$\log_2 B = \log_2 \left(2^k (1+x) \right)$$
(5)
= $k + \log_2(1+x)$ (6)

Al, Solarpunk, and an Uncertain Future in Computing Land Thank You

└─In Formal Terms, Cont.

In Formal Terms, Cont.	
- Let \mathbf{x} be: $\mathbf{x} = \sum_{i=0}^{k-1} 2^{i-k} b_i$	
In Formal Terms, Cont.

AI, Solarpunk, and an Uncertain Future in Computing 2024-07-12 └─Thank You

In Formal Terms. Cont.

In Formal Terms, Cont.

To calculate the approximate product, we sum two approximate log values and calculate the approximate antik

• To complete the calculation, we approximate. The mantissa $log_2(1+x)$ is approx. x, since $0 \le x < 1$. Thus:

$\log_2 B \approx k + x$ (7)

To calculate the approximate product, we sum two approximate log values and calculate the approximate antilog using a similar method.

Quantifying Area, Power, and Error

- For int16, MA is 70% smaller and uses 78% less power than accurate multipliers
- However, error is both large and unpredictable



- Mean relative error of 3.7%, peak relative error of 11.1%
- How can we do better?

AI, Solarpunk, and an Uncertain Future in Computing

Quantifying Area, Power, and Error



Quantifying Area, Power, and Error

Mean relative error of 3.7%, peak relative error of 11.1% How can we do better?

-07-12

2024-

Minimally Biased Multipliers

• MA + Novel error-reduction scheme

 Within each pair k₁ and k₂, error E differs by a scaling factor of 2^{k₁+k₂}:

$$egin{aligned} & ar{\mathcal{P}} &= egin{aligned} & ar{\mathcal{P}} &= egin{aligned} & -2^{k_1+k_2}(x_1x_2); & x_1+x_2 < 1 & (1) \ & -2^{k_1+k_2}(1+x_1x_2-x_1-x_2); & x_1+x_2 \geq 1 & \end{array} \end{aligned}$$

- The mean error within each pair is $-0.08333 \times 2^{k_1+k_2}$
- We can use this as an error-reduction term

AI, Solarpunk, and an Uncertain Future in Computing

Minimally Biased Multipliers

Minimally Biased Multipliers

• MA + Novel error-reduction scheme • With near sparin k_1 and k_0 , error E differs by a scaling factor $g(2^{k_1+i_1})$, $E_p=\bar{P}-\bar{P}$ $= \begin{cases} -2^{k_1+k_2}(x_1x_2) & x_1+x_2 < 1 \\ -2^{k_1+k_2}(x_1x_2-x_1-x_2), & x_1+x_2 > 1 \end{cases}$ • The mean error within each pair is $-0.0333 \times 2^{k_1+i_2}$

-07-12

2024

Quantifying MBM vs. MA

• MBMs are larger than MAs but much more accurate

int8	Error Bias	Peak Error	Area Red.	Power Red.
MBM	0.05	7.81	35.2	38.7
MA	-3.76	11.11	50.2	54.6

int16	Error Bias	Peak Error	Area Red.	Power Red.
MBM	-0.09	7.81	63.9	73.7
MA	-3.85	11.11	69.5	77.8

All values in %

- We can create float MBMs by replacing mantissa multiplication in IEE 754 with MBMs.
 - 57x power and 28x area improvement
 - 25% peak error, 4% error bias

AI, Solarpunk, and an Uncertain Future in Computing

└─Quantifying MBM vs. MA

Quantifying MBM vs. MA								
MBMs are larger than MAs but much more accurate								
	int8	Error Bias	Peak Error	Area Red.	Power Red.			
	MBM	0.05	7.81	35.2	38.7			
	MA	-3.76	11.11	50.2	54.6			
Content Processing I have been been been been								
	anr20	Error blas	Peak Error	Anea Red.	POWER PLEG.			
	MBW	-0.09	7.81	63.9	73.7			
	MA	-3.85	11.11	69.5	77.8			
All values in %								
 We can create float MBMs by replacing mantissa multiplication in IEE 754 with MBMs. 								

Mention error bias, all errors essentially average to near zero

-07-12

2024