AMPERE®

Balancing Intuition and Data in CPU Performance Modeling

Mahesh Madhav

Fellow, CPU Performance and Media Production

August 12, 2026

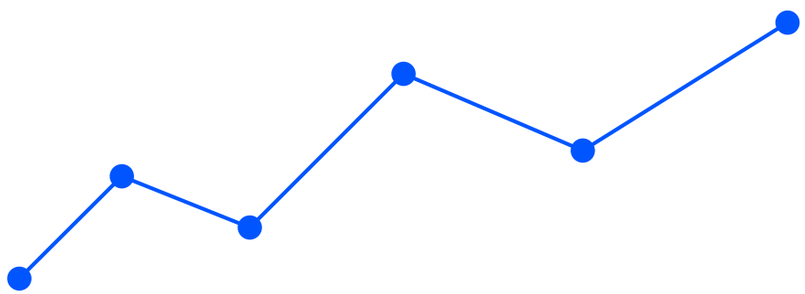
My Background

- Poking at CPU performance modeling and simulation since 1998
- Internships at **Digital Equipment Corp** and **Sun Microsystems**
- 16 years at **Intel Corp**
  - Willy, Coho, Keiko, Larrysim, HastyCore, LIT tracing
- 8 years at **Ampere Computing**
  - Panthera, QEMU tracing
- 4 years on **SPEC CPU Committee**
  - CPU2026 (Led 25 candidates → 15 accepted)
  - Rolling Round Robin
- Filmmaker and Media Producer
  - Directed 1 short doc on Amazon Prime, 1 to be released…
  - Produced 13 short films
  - Erdős – Bacon Number = 7

The two wings of a bird

DATA-DRIVEN ANALYSIS

- Simulation
- Regressions
- Telemetry



ARCHITECTURAL INTUITION

- Experience
- Judgment
- "Gut Feel"

Our story begins in the late 1990s...

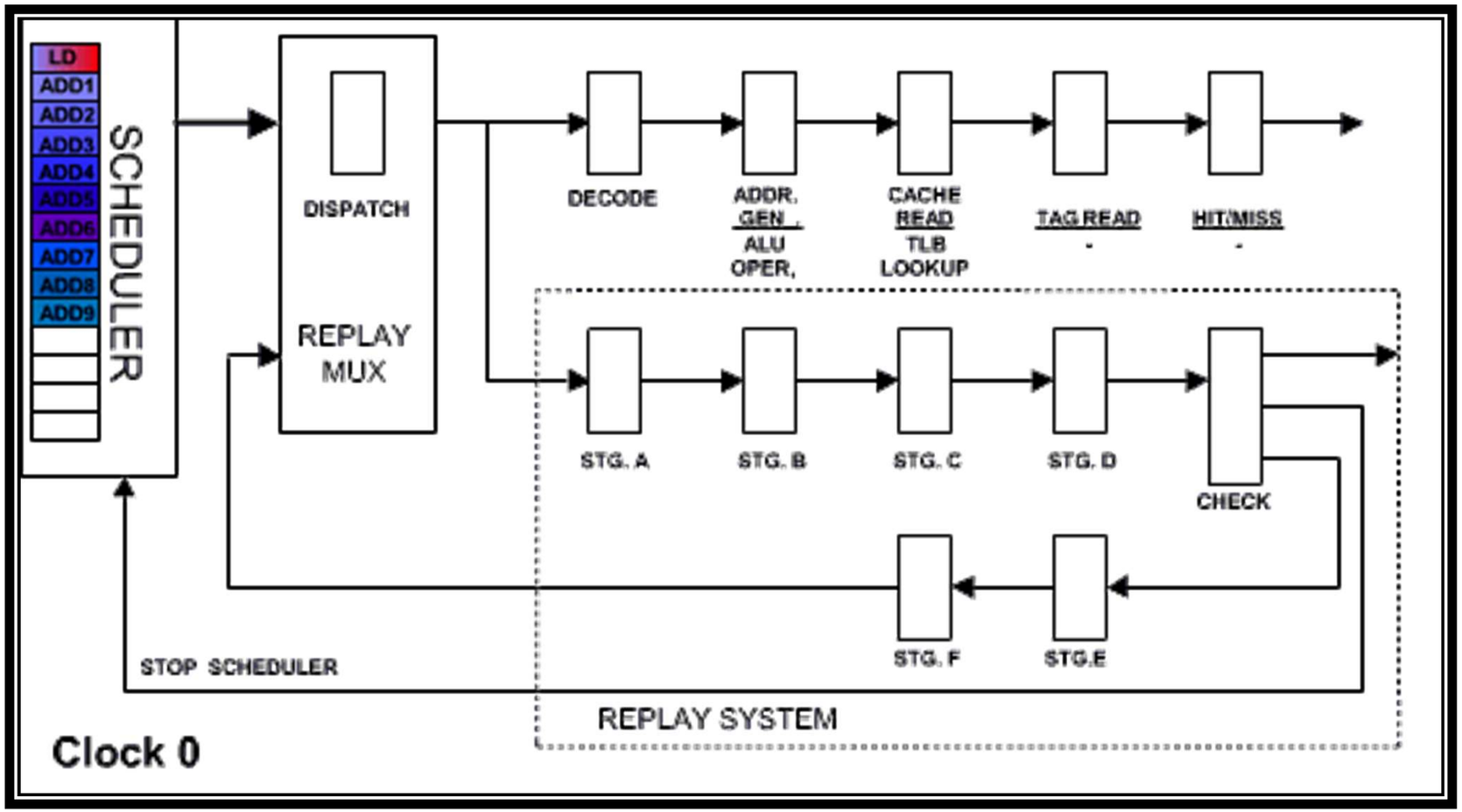
Intel Netburst:

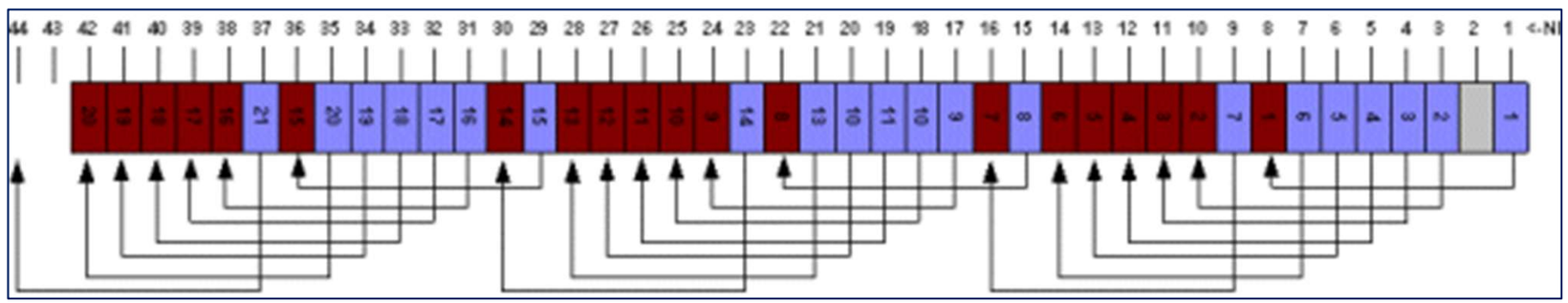
Pentium 4's Data Speculative uArch

Execution Replay is akin to "spin wait"

Replay causes include:
L1 cache miss, DTLB miss, missed STLF, variable latency FP ops

5 independent schedulers!
Replay tornadoes!
Livelock breakers!





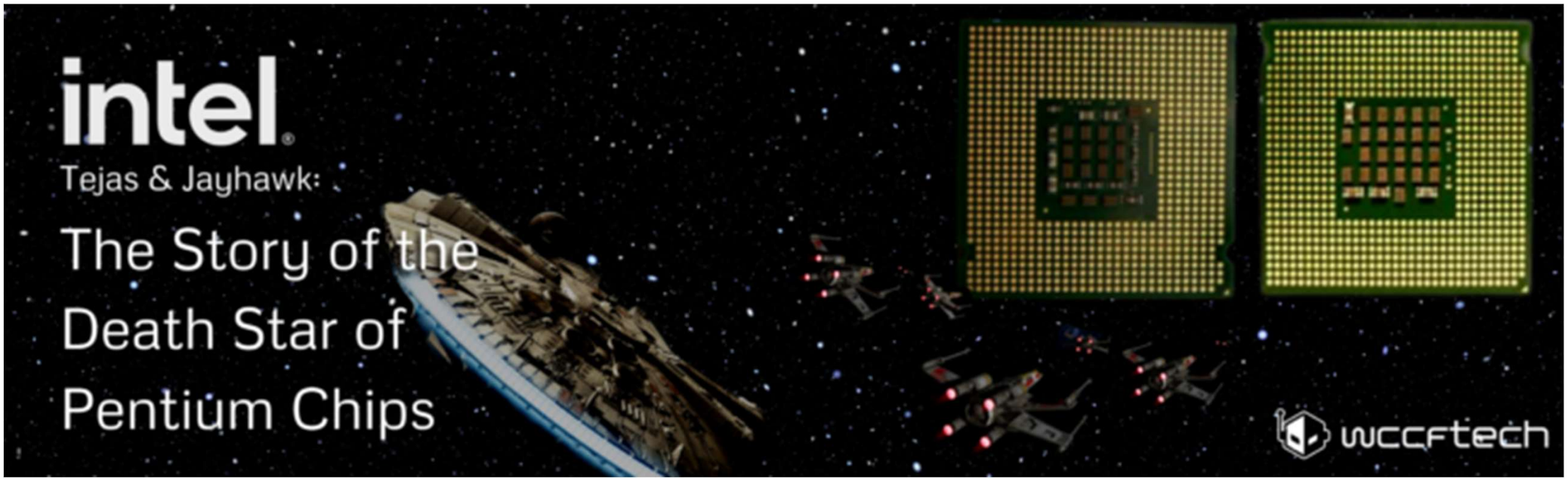
“The first **model** I designed was quite naturally perfect. It was a work of art. Flawless. Sublime. **A triumph only equaled by its monumental failure.**"

The Matrix Reloaded, 2003

The need for higher fidelity data (1999)

- First P4 models were DFA and NDFA queueing models.
  - Non-deterministic Finite Automaton
- These failed to predict the slow performance shown by the RTL, which led to the first Intel “Performance Dungeon”
- From this Dungeon came the exec@exec performance simulator
  - Willy (named for the Willamette River, the code name of the first P4 uarch)
- Very detailed model written in C
  - functionality coupled to performance execution
  - real data movement carried throughout core to memory
  - checkers at retirement to catch modeling errors.
- Thus, the model correlated with RTL!

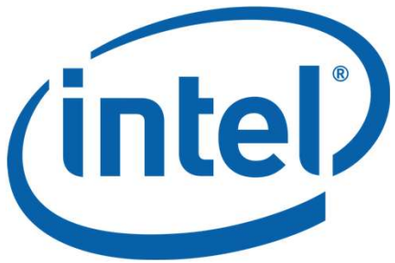
P4/Netburst gave way to P6/Nehalem uarch (2003)

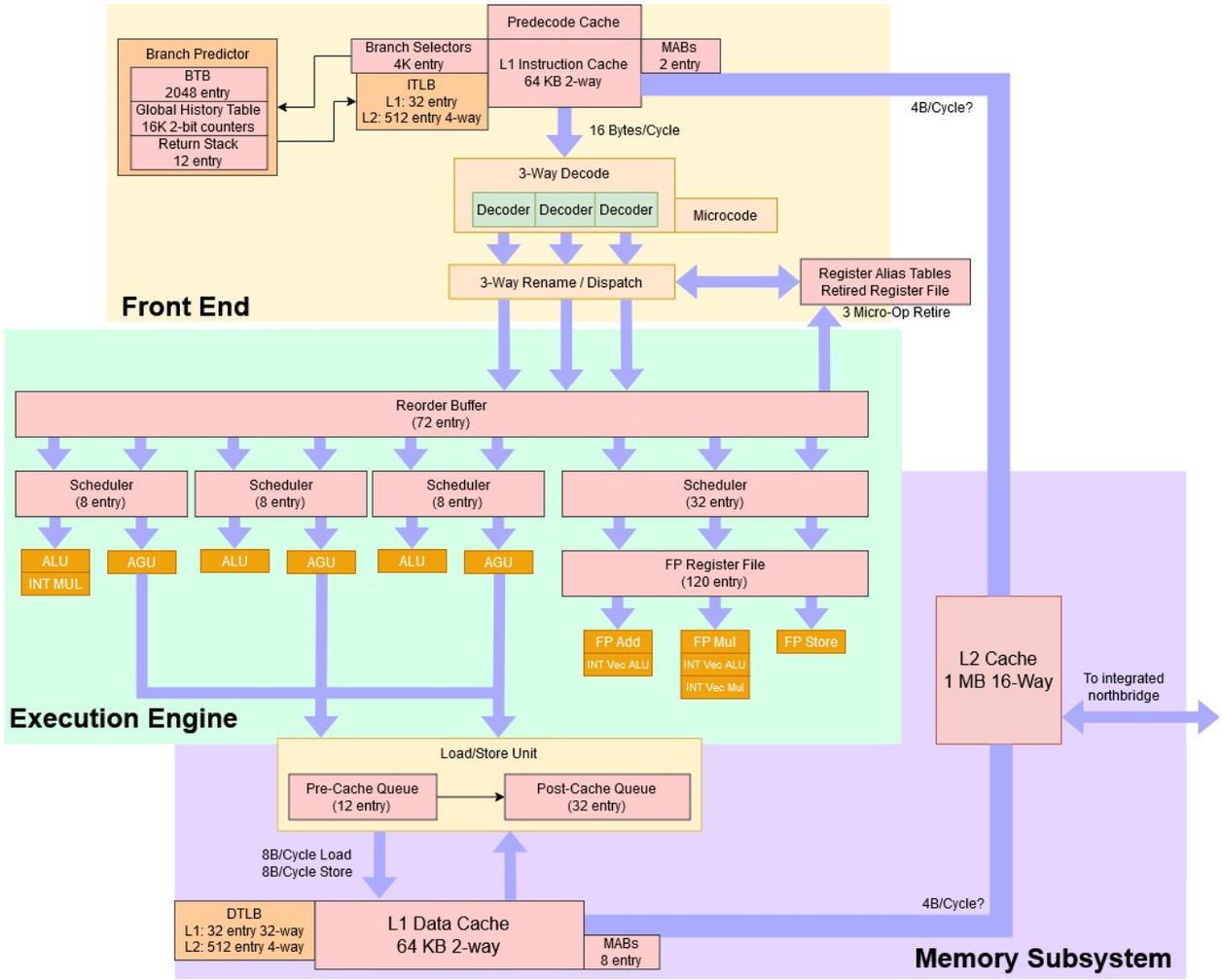


https://wccftech.com/intel-tejas-and-jayhawk-the-story-of-the-abandoned-intel-7-ghz-pentium-5-chips/

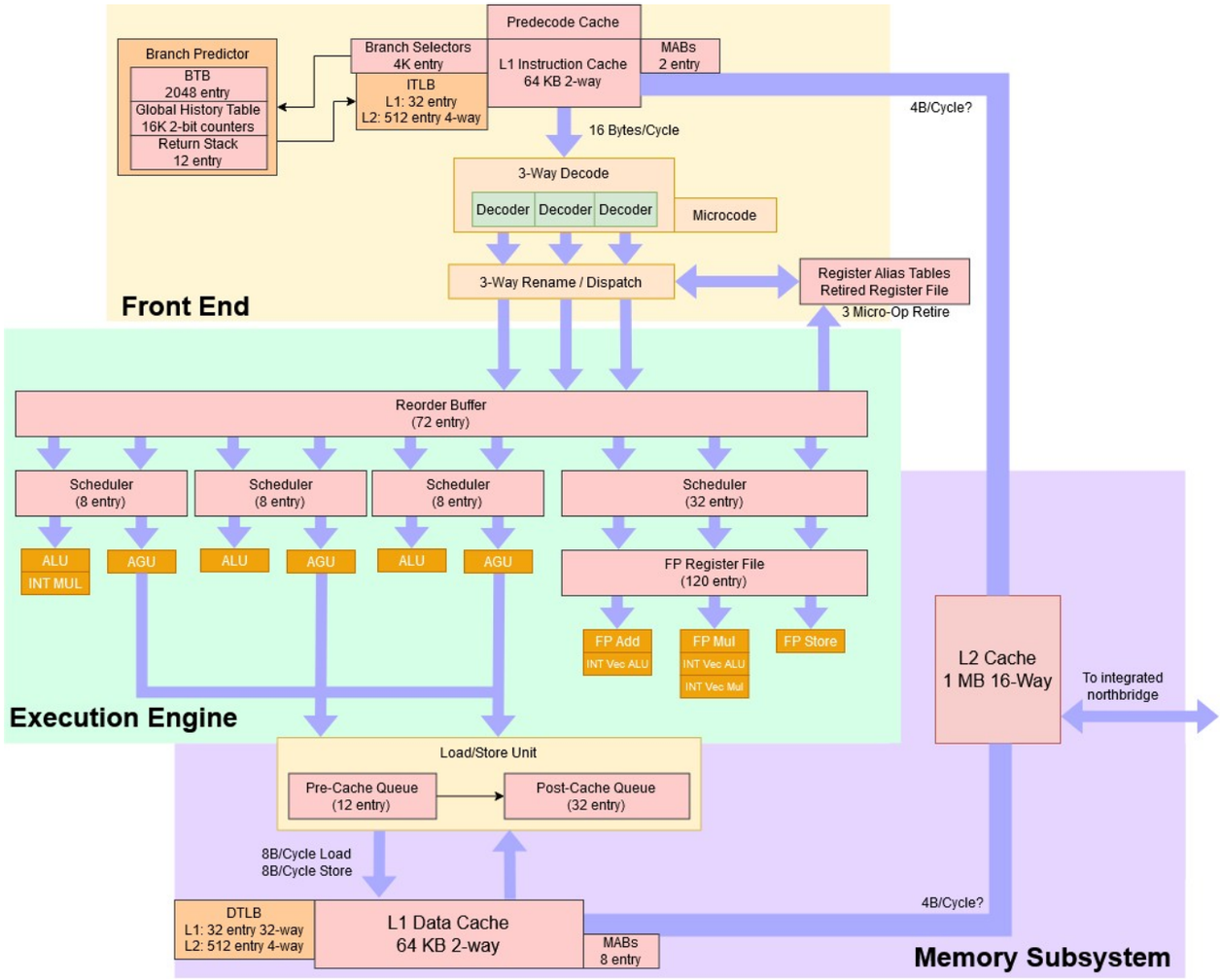
Coho/Keiko: A new performance model (2004)

- Very detailed model written in C++, motivated by P4 modeling issues
  - Built on the infrastructure started by Willy, plus:
  - Object Oriented Modeling, cycle accuracy
  - Data Checkers to prevent cheating in the model
  - Tremendous amount of telemetry for debugging

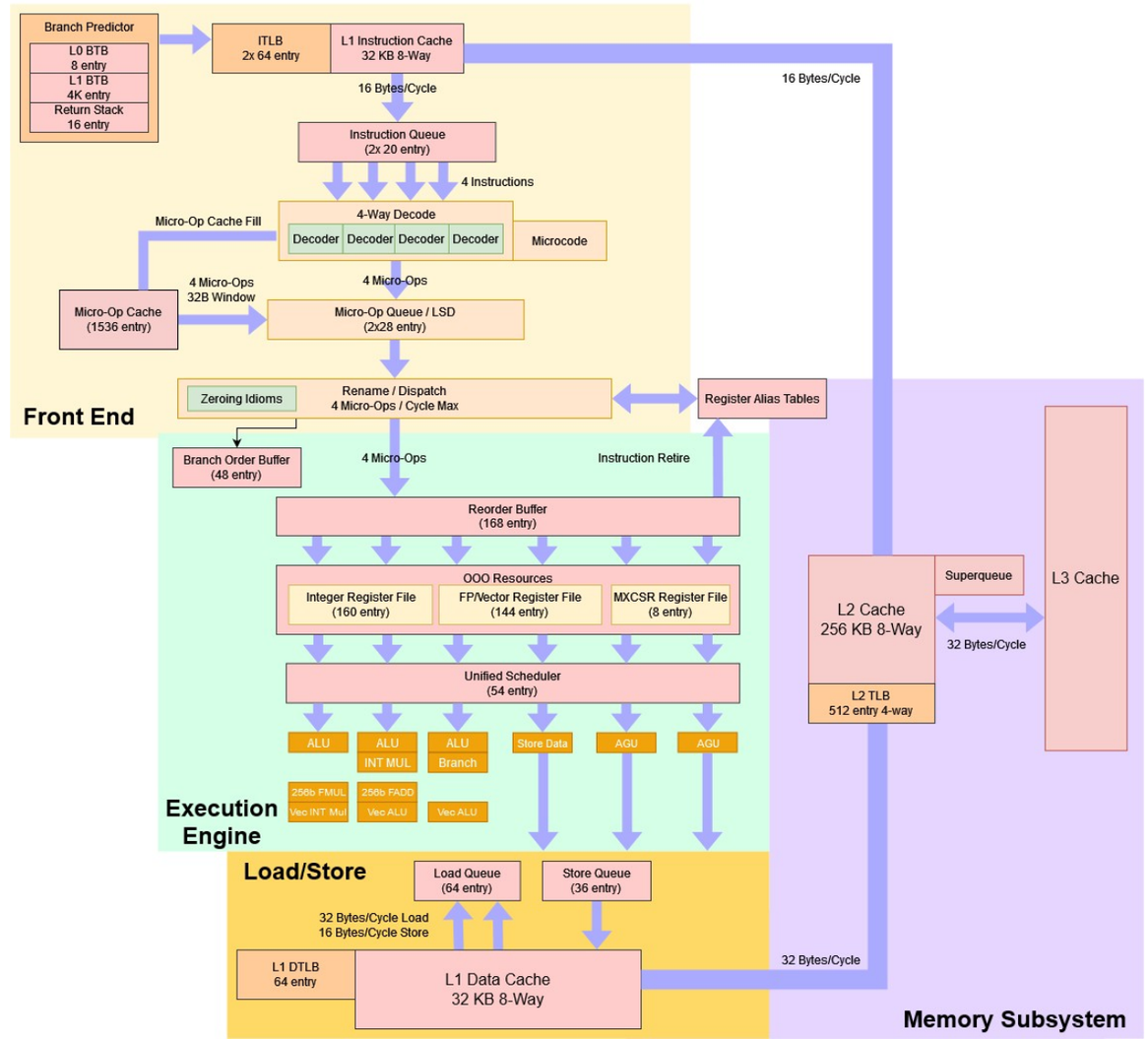




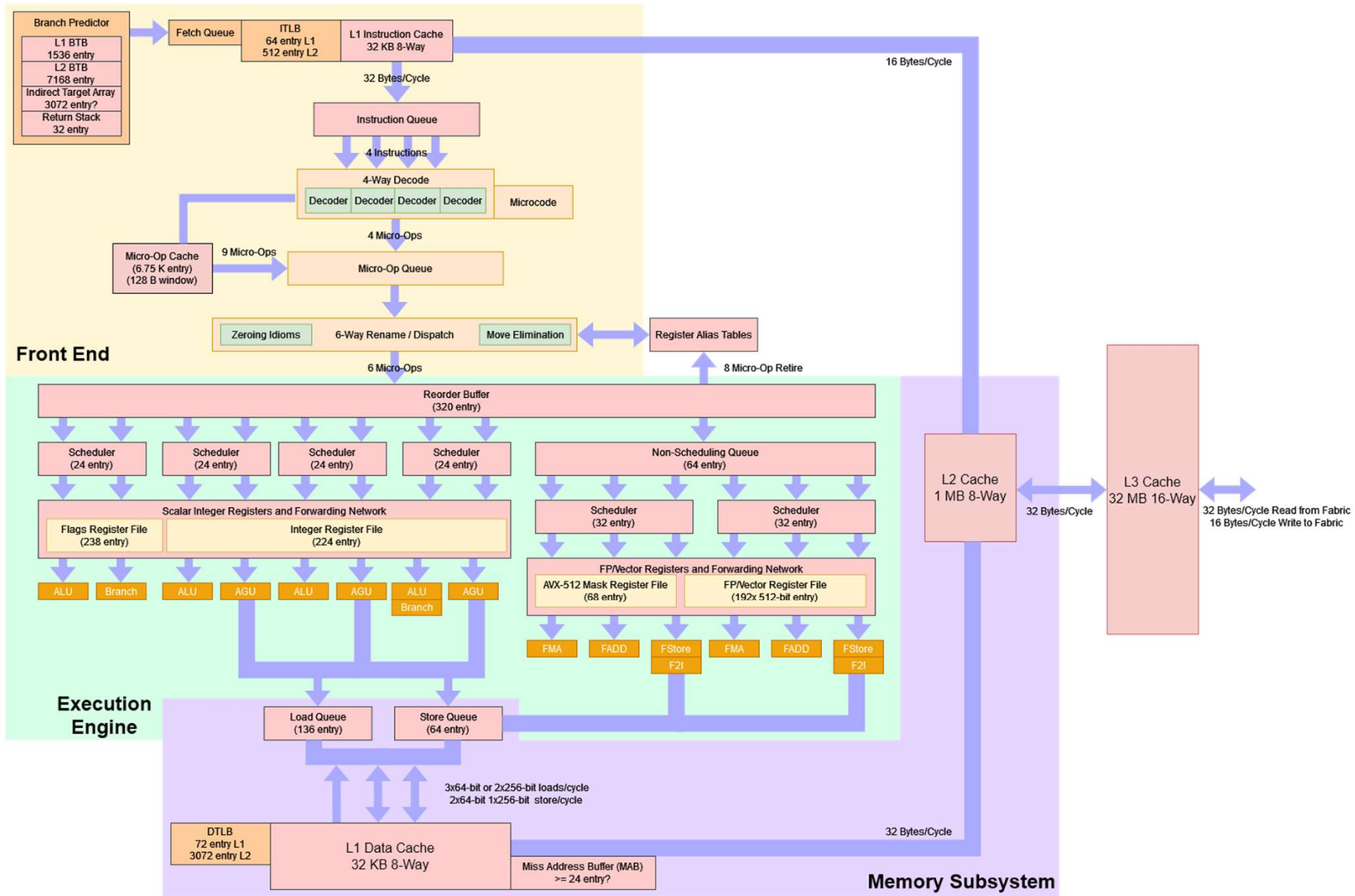
Intel Sandy Bridge (2011)



AMD Bulldozer (2011)



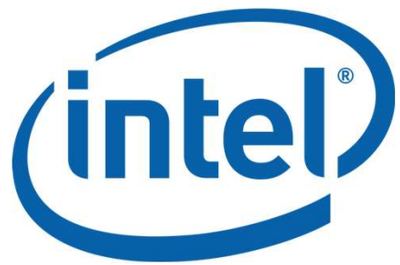
Intel Sandy Bridge (2011)



AMD Zen4 (2023)

Coho/Keiko: A new performance model (2004)

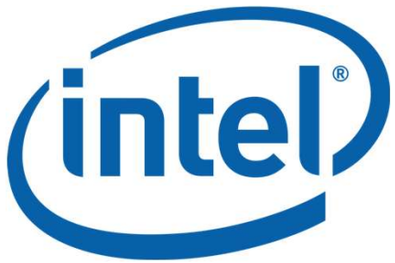
- Very detailed model written in C++, motivated by P4 modeling issues
  - Built on the infrastructure started by Willy, plus:
  - Object Oriented Modeling, cycle accuracy
  - Data Checkers to prevent cheating in the model
  - Tremendous amount of telemetry for debugging
- The details were unnecessary because the uArch was simpler
  - Design space exploration became much harder
  - Complex and exact modeling required for all pathfinding features
  - Certain knobs began to rust in place
  - Oracle studies were almost impossible
  - Twenty years later, Intel is still using this framework in production



Loss of intuition

- Narrow explorations led to atrophy of intuition
- Junior architects became bound by simulator study outputs
  - Study volume increased but understanding did not

- Maybe this was an economic issue…



“I redesigned it… However I was again frustrated by failure. I have since come to understand that the answer eluded me because it required a mind less bound by the parameters of perfection. **Thus, the answer was stumbled upon by another, an intuitive program.**”

The Matrix Reloaded, 2003

Panthera: A new performance model (2018)

- Learning from the mistakes of our past
- Guided by the constraints of an agile startup company
- Built to the needs of the nimble arch team **designing a new core**
  - exec@fetch model, trace driven with warmup
  - no wrong-path modeling
  - no functional simulator, only a page translation checker
  - no multi-threading, no multi-core
- Traces for perf model and RTL had to be collected simultaneously
  - RTL checkpoints are crafted from ELF memory dumps…
  - While sim traces are just a list of instructions to be replayed

AMPERE®

Gaining intuition

- Many early uarch choices were made through intuition and experience
- Process of model bring-up and performance verification led to new uarch ideas for next generations.
- Pretty good correlation between performance model and first silicon

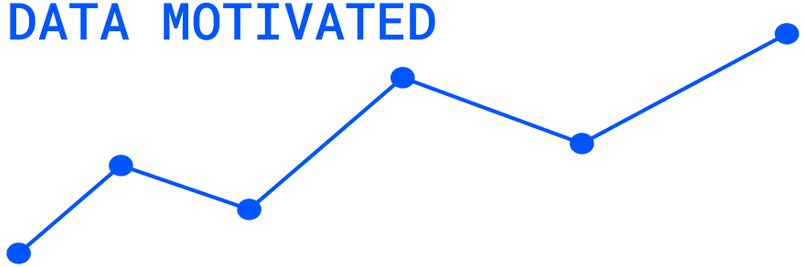
CPU2017 benchmark	96 cores	128 cores	192 cores
500.perlbench_r	0.99	0.98	0.98
502.gcc_r	1.06	1.05	1.05
505.mcf_r	0.88	0.90	1.03
520.omnetpp_r	1.04	1.06	1.01
523.xalancbmk_r	0.84	0.82	0.80
525.x264_r	0.99	0.99	0.99
531.deepsjeng_r	1.06	1.06	1.08
541.leela_r	0.99	0.98	0.97
548.exchange2_r	1.02	1.02	1.02
557.xz_r	0.91	0.92	0.93

TABLE I
BASELINE SPECRATE CORRELATION FOR AMPEREONE SOCS.

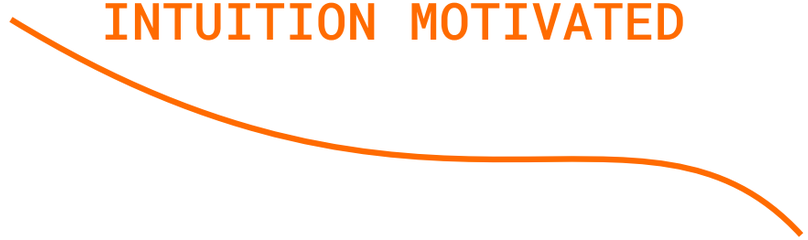
https://arxiv.org/abs/2506.02344

AMPERE®

DATA MOTIVATED



INTUITION MOTIVATED



NDFA

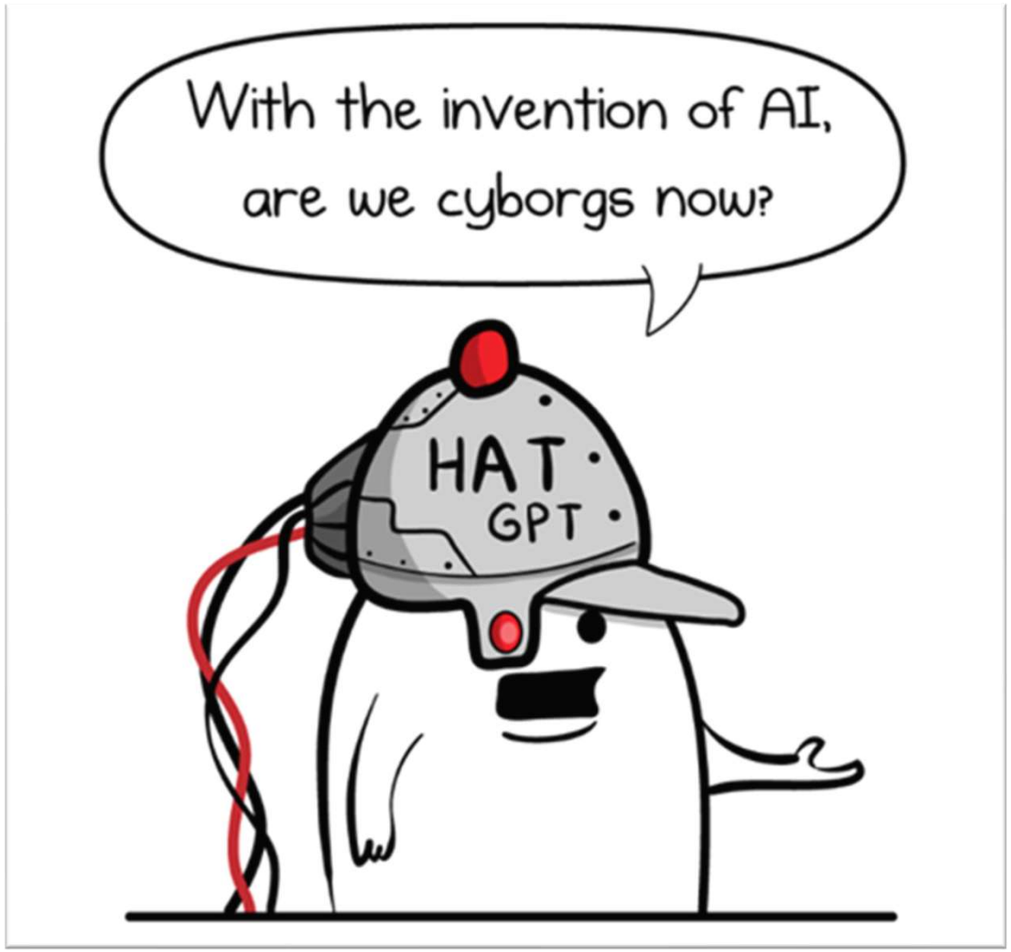
Willy

Coho

Panthera

But that all happened before 2024...

Acknowledgement of symbiosis

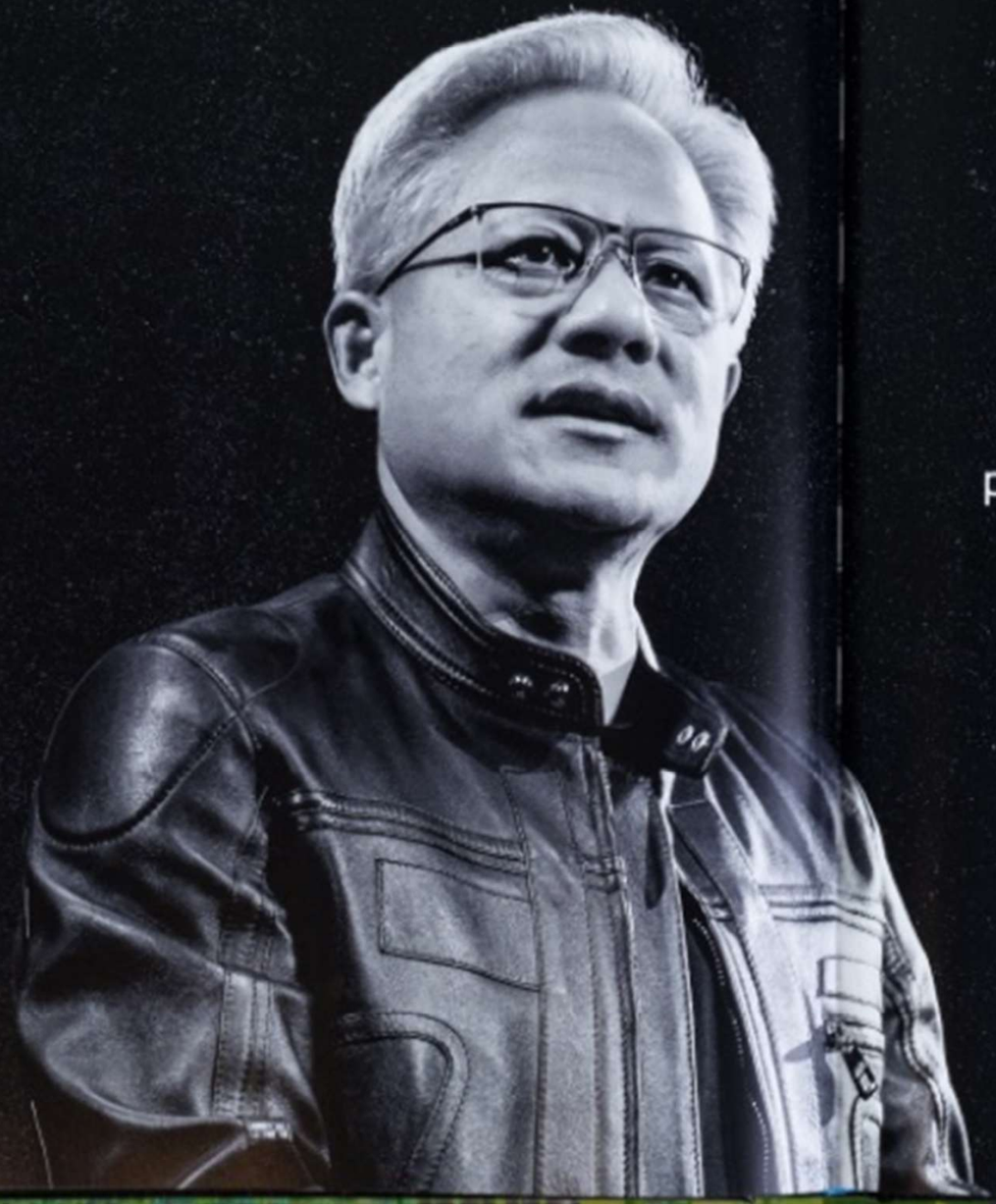


Hard work turns into intuition

Intuition requires constant exercise

Danger arises when LLM access is mistaken for expertise

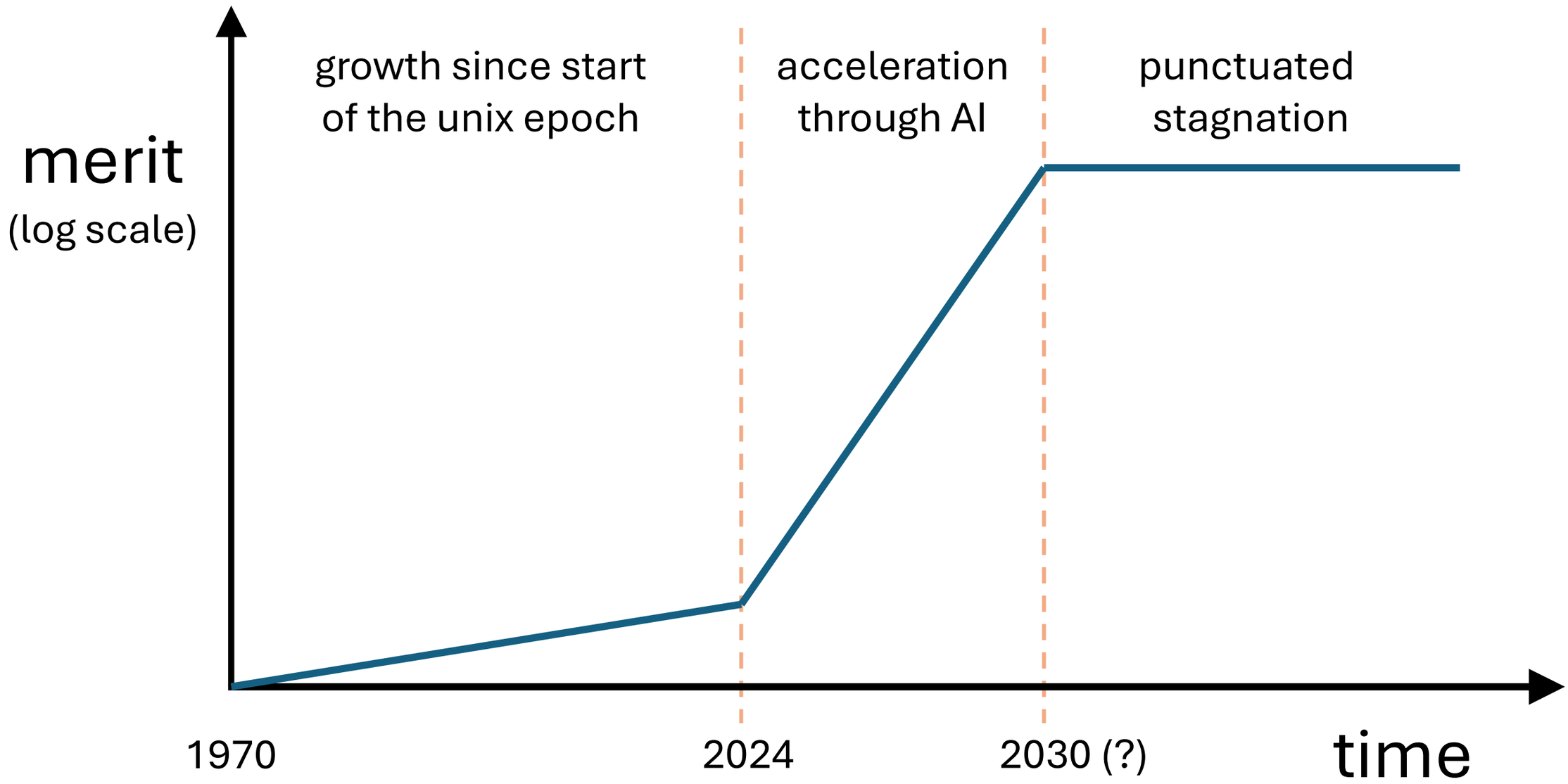
October 2025
https://theoatmeal.com/comics/ai_art



“Greatness comes from character. And character isn’t formed out of smart people, it’s formed out of people who suffered.

I wish upon you ample doses of pain and suffering.”

March 2024


merit
(log scale)
growth since start
of the unix epoch
acceleration
through AI
punctuated
stagnation
1970
2024
2030 (?)
time

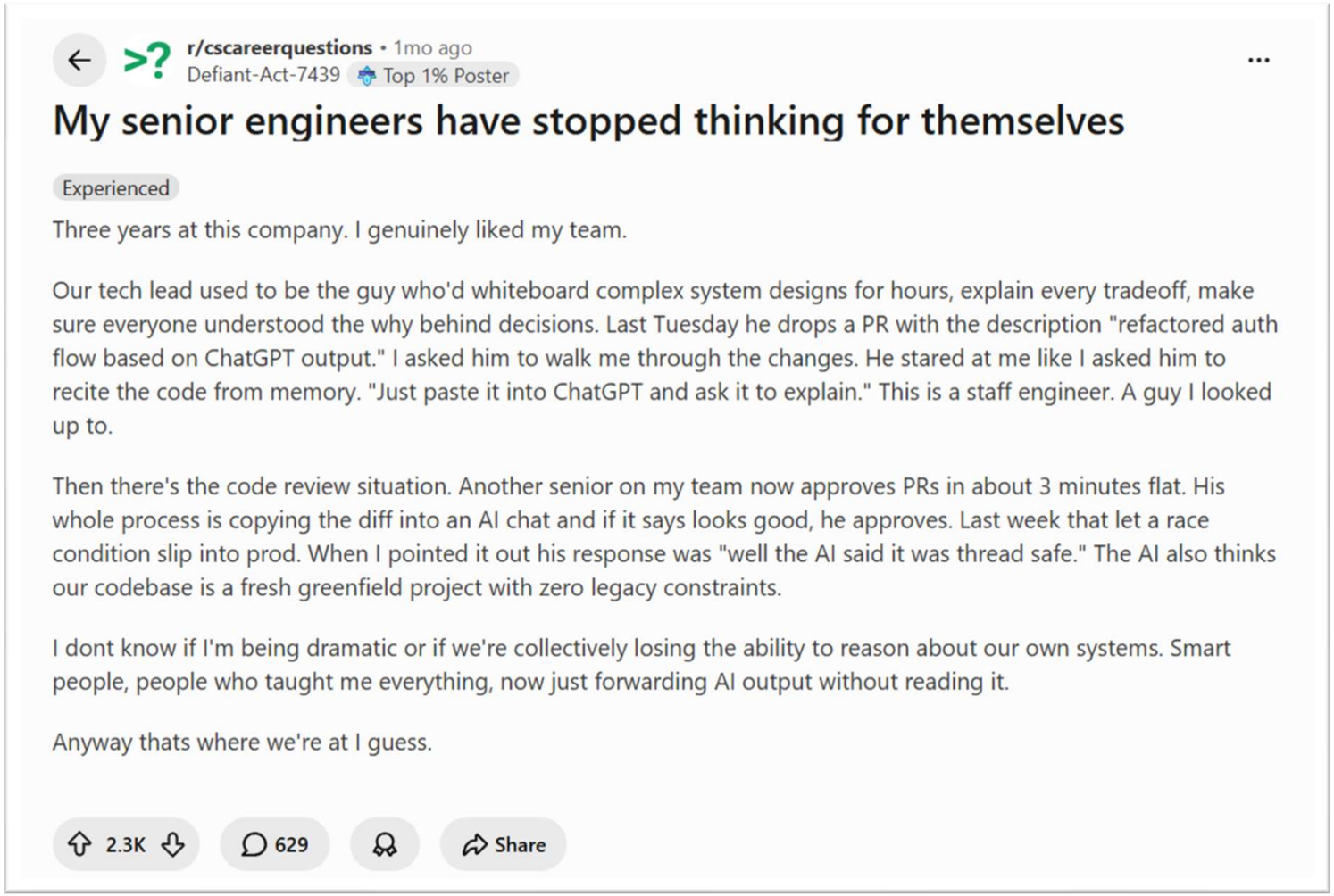
Race to halt!

If the “merit” takes a long time to reach a plateau …

- Students don’t enroll in CompArch
- Junior Architects shift industries
- Senior Architects retire

… Computer Architects go extinct.

We are turning off our brains ...



r/cscareerquestions • 1mo ago
Defiant-Act-7439 Top 1% Poster

My senior engineers have stopped thinking for themselves

Experienced

Three years at this company. I genuinely liked my team.

Our tech lead used to be the guy who'd whiteboard complex system designs for hours, explain every tradeoff, make sure everyone understood the why behind decisions. Last Tuesday he drops a PR with the description "refactored auth flow based on ChatGPT output." I asked him to walk me through the changes. He stared at me like I asked him to recite the code from memory. "Just paste it into ChatGPT and ask it to explain." This is a staff engineer. A guy I looked up to.

Then there's the code review situation. Another senior on my team now approves PRs in about 3 minutes flat. His whole process is copying the diff into an AI chat and if it says looks good, he approves. Last week that let a race condition slip into prod. When I pointed it out his response was "well the AI said it was thread safe." The AI also thinks our codebase is a fresh greenfield project with zero legacy constraints.

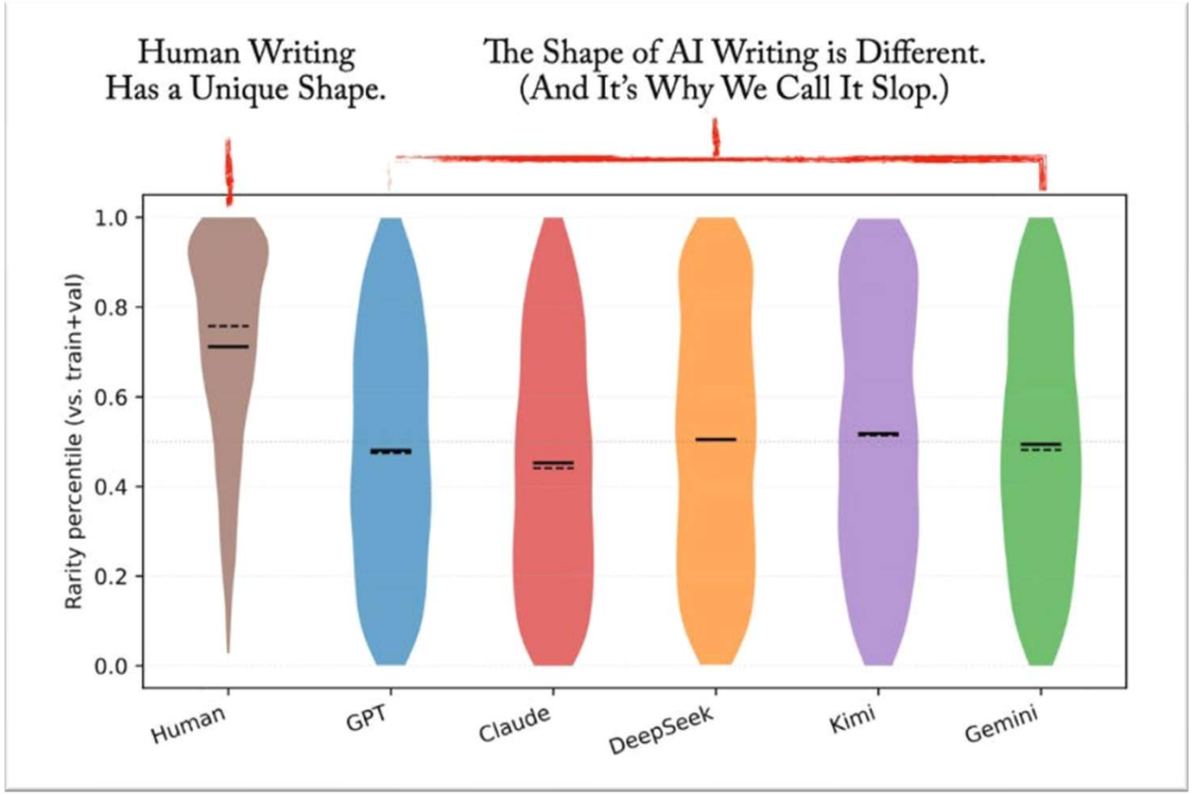
I dont know if I'm being dramatic or if we're collectively losing the ability to reason about our own systems. Smart people, people who taught me everything, now just forwarding AI output without reading it.

Anyway thats where we're at I guess.

2.3K 629 Share

May 2026, Reddit
https://www.reddit.com/r/cscareerquestions/comments/1tkccsi/my_senior_engineers_have_stopped_thinking_for/

Regression to a terrible mean ...



StoryScope: Investigating idiosyncrasies in AI fiction
April 2026, U-Maryland and Google
https://arxiv.org/abs/2604.03136

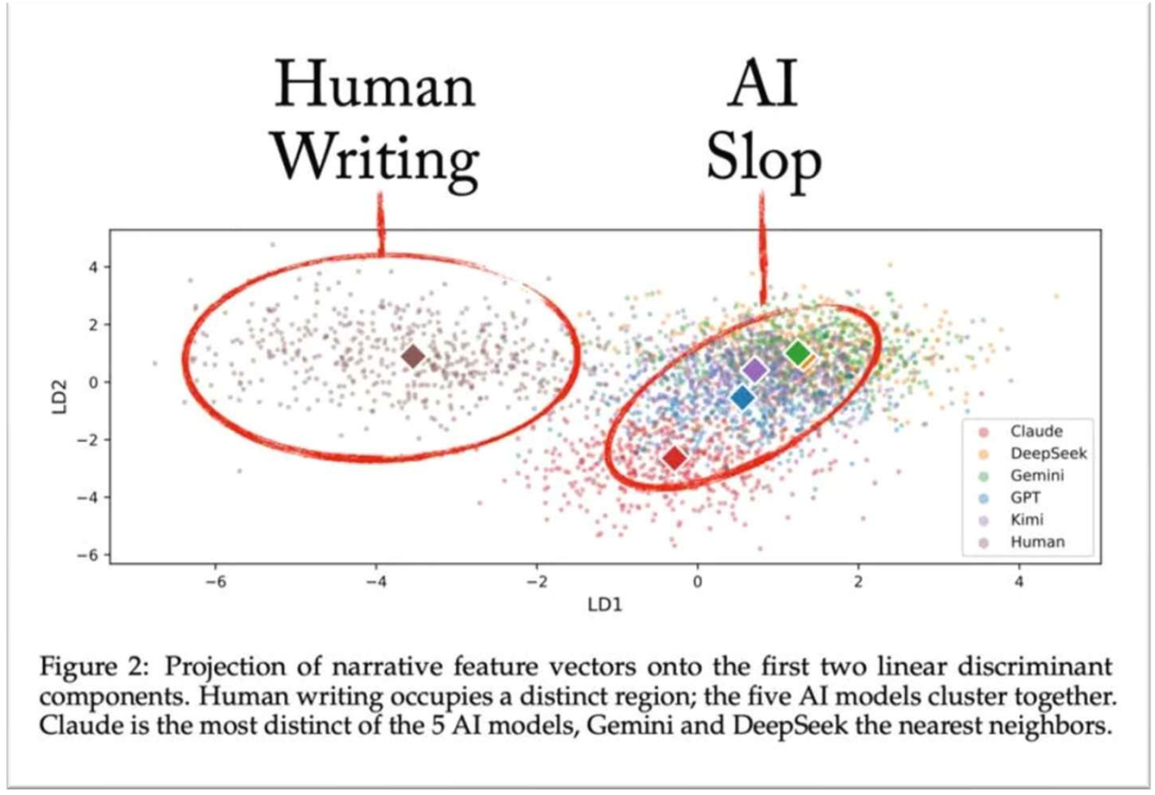


Figure 2: Projection of narrative feature vectors onto the first two linear discriminant components. Human writing occupies a distinct region; the five AI models cluster together. Claude is the most distinct of the 5 AI models, Gemini and DeepSeek the nearest neighbors.

The Shape of Enshittification, June 2026
https://thedigitalcontrarian.substack.com/p/the-shape-of-enshittification-104

Model Misalignment and Language Change: Traces of AI-Associated Language in Unscripted Spoken English

Bryce Anderson*, **Riley Galpin**, **Tom S. Juzek***

Florida State University

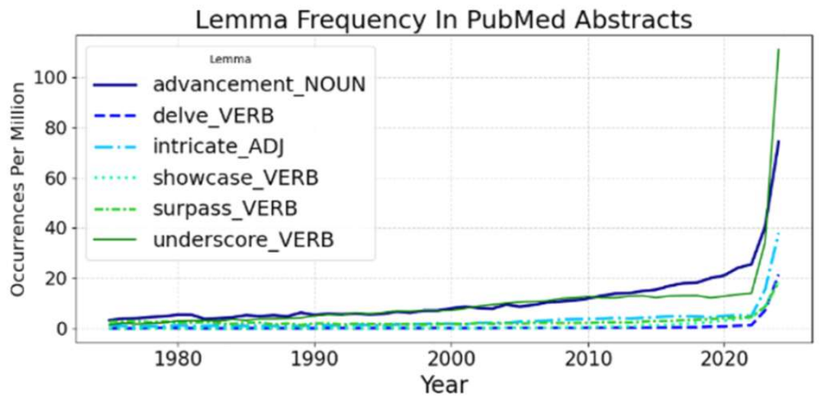


Figure 1: Whether the sharp post-2022 rise in AI-associated words reflects mere tool usage or a direct influence on the human language system remains an open question with broader societal relevance.

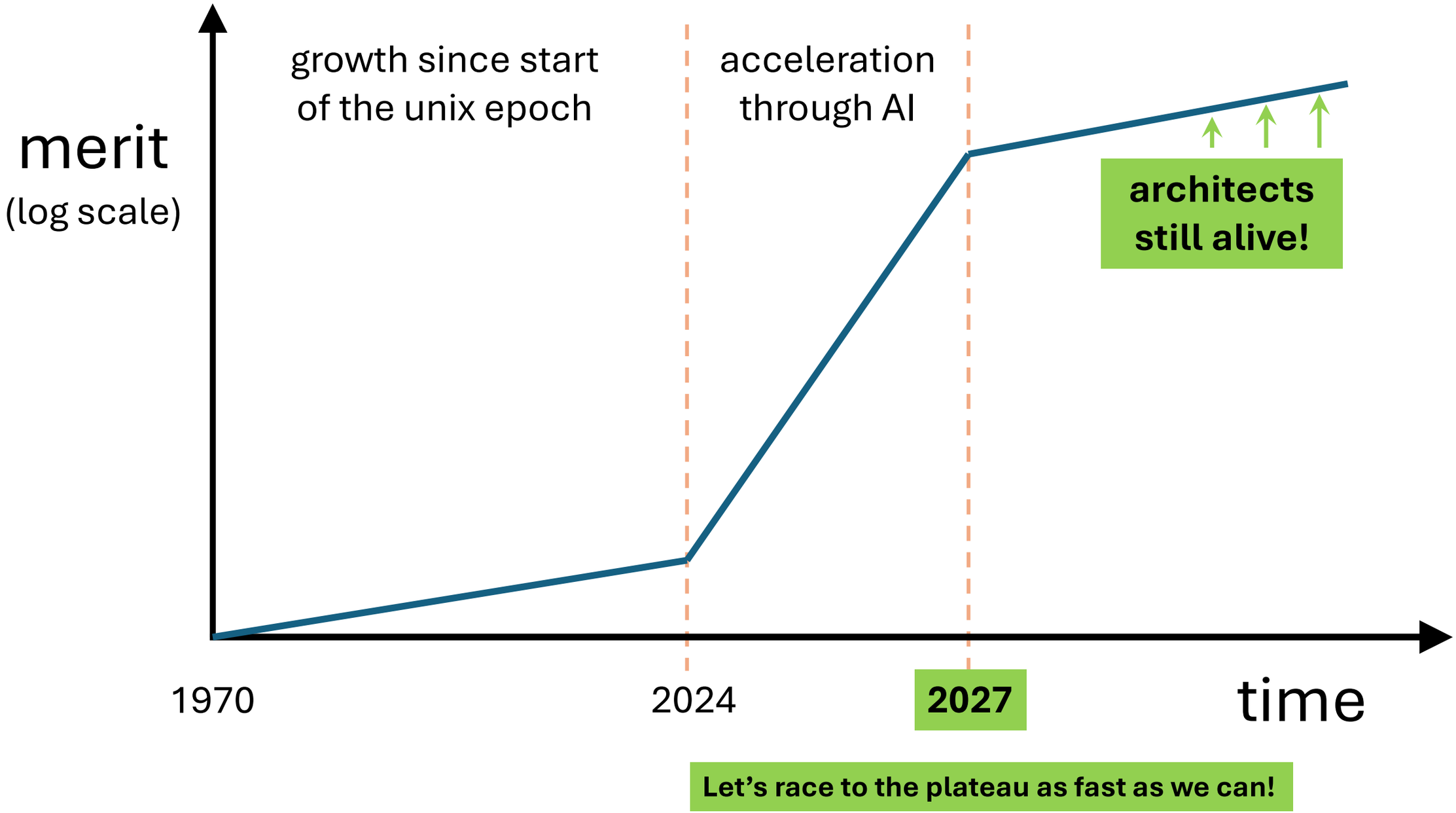
Words that have recently spiked in human speech:

notable, surpass, boast, meticulous, garner,

capability, assemble, reinforce, novel, integrate,

key, progress, crucial, feasible, domain

https://doi.org/10.1609/aies.v8i1.36540


merit
(log scale)
growth since start
of the unix epoch
acceleration
through AI
architects
still alive!
1970
2024
2027
time
Let's race to the plateau as fast as we can!

Reflections

Don’t blindly rely on the output of AI
Be a pilot, not a passenger

Mentor juniors through the struggle
Make them earn their intuition (through suffering)

Active guidance: keep humans in the loop
Remain the grand central station of intuition

Demand the “why”
Understanding the root cause is a requirement

Race to the plateau
Use AI aggressively to find its limits quickly

You make meaning through the doing
Ideas are common, effort is not: no shortcuts

go fast!

Thank You

Mahesh Madhav

mahesh@amperecomputing.com