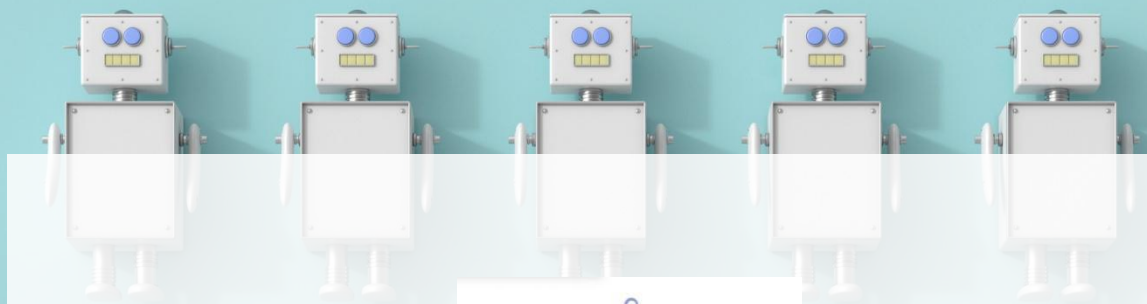


ACHIEVE

Co-funded by the European Union

Affordability in AI Computing

Trends and troubles

Authors: Veljko Petrović
Institution: University of Novi Sad Faculty of Technical Sciences

Affordability in AI Computing

Trends and troubles

Author: Veljko Petrović

Institution: University of Novi Sad Faculty of Technical Sciences

Abstract: Effective education in AI and AI computing requires yielding to what Steven Levy called "the hands-on imperative." Prospective students, on their journey to mastery, must be able to get their hands on the equipment in question, and have the opportunity to 'waste time' on it, experimenting, exploring, and pursuing quixotic goals if maximum results are to be achieved. Given the prices involved in AI computing, this is difficult to achieve. A number of options are considered for a budget-conscious educational program, and suggestions are offered as to an optimal mix of technologies and approaches.

Wasting (Computer) Time

There is nothing quite so conductive to learning a field of computing as wasting time. Certainly, structured learning has its place, and it is indispensable to promoting a healthy understanding of the fundamentals, but when it comes to truly internalizing how something works, few things come close to the deceptively useful pastime of messing about.

Over the years, people have learned various languages, frameworks, and programming approaches by building toys, or deliberately inefficient and labor-intensive solutions to everyday problems. This goes back to the very earliest days of computers, when the proto-hackers cut their teeth on ultra-expensive and exotic TX-0 and PDP-1 machines building such wastes of time as the Expensive Desk Calculator (what is likely the first interactive computer calculator program), the Expensive Typewriter (one of the first word processors), and of course the very worst of all student excesses…

It would be *remarkably* difficult to defend the use of countless hours of, at the time, incredibly precious computer time (A PDP-1's cost is something like $1,300,000 in today's dollars) to a grant

ACHIEVE project is developed and delivered under European Union's Digital Europe Programme Project no. 101190015

Co-funded by the European Union

committee in the service of the creation of what must have been, at the time, the world's most grotesquely overpriced toy.

And yet, people who did all that messing about and time-wasting are in large part responsible for the computing world we have today. Indeed, the very term 'Artificial Intelligence' now so important, came about during those early heady days at MIT.

Spacewar! Not the first computer game, but remarkably close.
(Image, "Spacewar running on PDP-1" courtesy of Joi Ito from Inbamura, Japan, CC-BY-2.0

Of course, the students there who had largely unfettered access to those early machines were quite fortunate indeed. It wouldn't be until the eighties until it was possible for most anyone to get their hands on a computer all of their own to waste time on without limits. This prolific period of time-wasting brought us the personal computing revolution we are still coping with, decades later.

This all, naturally, brings us to today.

ACHIEVE project is developed and delivered under European Union's Digital Europe Programme Project no. 101190015

Co-funded by the European Union

What's old is new again

We are, in a very real sense, in a repeat of the 1970s when it comes to AI. For decades we have enjoyed the fruits of the personal computing revolution which allowed for near-unlimited time-wasting, experimentation, and vast amounts of what Richard Feynman, it is said, called 'active irresponsibility.' Computers aren't cheap, but just about anyone motivated enough could get enough computing power to do nearly anything that could be done on a computer. It was quite possible for a curious teen to pick up kernel development, and, thanks to the open source movement, contribute meaningfully from a bedroom computer. We have all been the beneficiaries of the development of this remarkable state of affairs.

AI broke that model. To do truly interesting work in this crucial field, one needs access to grades of computation ability that are simply inaccessible to a motivated individual. This is not that much of a barrier for production or for serious work: there is a surfeit of services that will rent out the necessary infrastructure at competitive rates. Serious work has not suffered, but messing about, on the other hand, is in crisis. Time on the big GPUs is charged by the second, just like it was on the mainframes of yesteryear, and it is quite difficult to justify the expenditure when the goal is messing about just to see what happens.

This will have knock-on effects both in development of the field (think, after all, how much of today's computing landscape we owe to motivated eccentrics working outside formal channels), and in the education of new experts.

What is required, therefore, is some way of kickstarting the 80s and 90s era of cheap universal computing, but for the AI era. A technological stack that can be made ubiquitous enough that there's room for messing about, and powerful enough that there's cause to mess about with it.

A wishlist for a ubiquitous AI workstation

The limiting factors here are cost, infrastructure requirements, and most crucially VRAM size. Speed is something that can be lived with: if training takes four times as long, or if inference is slow, this is something that patience can deal with in most use cases. The other requirements are less flexible.

ACHIEVE project is developed and delivered under European Union's Digital Europe Programme Project no. 101190015

Co-funded by the European Union

VRAM size is the biggest purely technical limiting factor: the size of VRAM is a hard limit on model complexity. If the model will not fit into VRAM, then no amount of patience will help. It is possible to cheat a bit by using aggressive quantization and accepting reduced performance in trade for a smaller model. However, quantization can only go so far: a modern model is as likely as not in FP16 or Bfloat16 to begin with and that represents a hard limit on how much you can claw back with quantization.

For making toy models, certainly, small VRAM sizes are not that much of a problem. But for *quality* messing about, it isn't about classifying MNIST so much anymore, as it is about building a tame LLM of your own and seeing what makes these things tick.

Cost and infrastructure go together as limitations: if the hardware is unaffordable or if it has cooling, power, or space requirements that are untenable, it doesn't matter how good it is. It simply can't achieve the necessary ubiquity.

What is required is a sort of Commodore 64 of the AI era: just enough computer to be dangerous, while still being available enough that it is not beyond imagining to have one on your desk. We aren't there yet, not *quite,* but it is remarkably close.

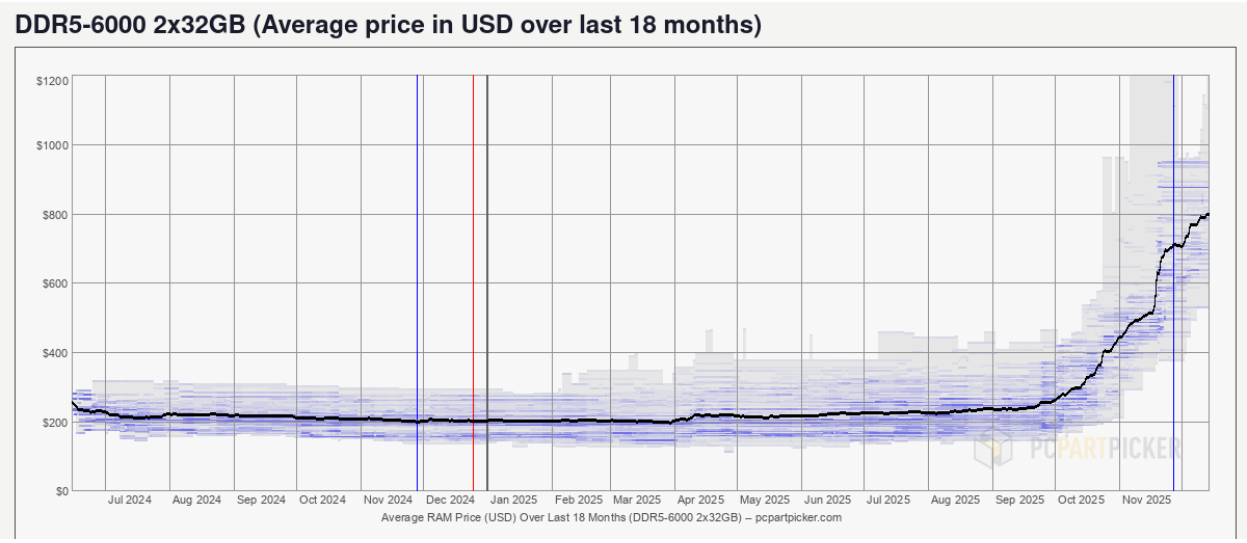
A roll-call of affordable hardware

The golden standard for affordable AI hardware is still the **gaming GPU**. These are at least *somewhat* affordable, due to economies of scale, and they represent the most TOPS-per-dollar that you can get. The software support, too, is excellent, as the major producers of these chips use fundamentally the same architecture on their supercomputing-grade and their consumer offerings which means that anyone using this approach can rely on best-in-class software support.

The chief issue is VRAM size. There isn't that much call to expand VRAM overmuch for a device whose ostensible use is still to push polygons in a video game. Even top-of-the-line models with exceptionally steep price points are limited to no more than 32 GB, and the ones that are actually worth the term 'affordable' will have considerably less.

A good trade can be to switch performance for memory: it is actually possible to get decent inference, at least, on **CPUs** with models that are state of the art, by taking a server-grade CPU

and speccing the computer with excess DRAM. You won't get large inference speeds, but you will get to experiment with models that you normally can't touch without a veritable *fortune* in hardware. And fortunately, at least DRAM is cheap…



Pictured: Murphy's law.
Source: https://pcpartpicker.com/trends/price/memory/

An attempt to square the circle and get plentiful RAM with at least some GPU performance is the **AI-enabled CPU.** It may be tempting to consider a Lunar Lake, Ryzen AI, or Snapdragon X Elite as a panacea in this situation, but seen in TOPS they trail some 2500-10000% behind top-of-the-line consumer GPUs. The ecosystem needed to leverage their NPUs is also still in its infancy and nothing like the robust toolsets that are available for uses of mainstream GPUs is available.

A **SoC with AI additions** will at least provide a unified and fast memory, and Apple Silicon would be a great choice if it weren't for the fact that it is completely locked down, and that it boasts a remarkably high price. Considerably more downmarket, is the **Raspberry Pi 5** equipped with an AI 'hat' like the **Hailo-8.**

This is in no way a powerful configuration, but it is remarkably affordable. Prices vary, but for about $300 it is possible to get 26 TOPS and 16 GB of unified memory in a complete system that

just needs a display and input devices to work. This is probably the cheapest way to get something that is 'messing about' ready available. A more commercial version of the same fundamental idea is the **Nvidia Jetson** series which provides respectable performance, though at a steep price when it comes to models with more expansive memory capabilities.

The most remarkable new development on this front is what may just be the Commodore 64 of the AI era, the 'shoebox supercomputers' that run on the **Nvidia DGX Spark** architecture. The price is not negligible, of course, but it provides about as much power as a serious workstation equipped with a 4090-grade GPU at *less* cost, and with a unified 128 GB memory which, NVidia claims is enough for 200B parameter models. The Spark is perhaps a bit too pricey for the average bedroom just yet, but it is eminently affordable as a time-shared resource to afford some *serious* messing about time, especially since its cooling and power requirements compare favorably to an office-grade desktop, let alone an AI workstation.

A plan for a new 80s

When it comes to budget-conscious education, the key is to allow as many students as much time to experiment with no end-goal in mind as possible. Getting the equivalent of a DGX Spark or unlimited supercomputer time is not practical politics for most places. The best approach, therefore, is split efforts between performance and ubiquity.

On one hand Raspberry PIs with AI hats can be obtained in large enough quantities that anyone really interested can be lent one for projects. A maximum spec Raspberry Pi with an AI hat and one of the better Pi cameras is a dream configuration for real-time, real-world, robotics-adjacent computer vision work, and it is portable enough to be deployed swiftly and anywhere an interesting problem may be found.

On the other hand, as many Nvidia DGX Spark or equivalent devices as can be afforded should be bought and, and this is crucial, *made available to interested students.* There is a tendency to treat expensive delights like these as crown jewels that need to be kept under lock and key and guarded by oversight, bureaucracy, and some sort of prioritization system. It is important to resist this urge: Put as few barriers between the machines and those interested, and time-share by time, not by task or officially sanctioned project. Give interested students the chance to 'waste' time on these machines, encourage curiosity and experiment, and–history tells us–the results are going to be *remarkable*.