Running machine learning models locally

▶ Table of contents

Introduction

This is a collection of links and information to assess the viability of running machine learning models and large language models locally. The aim is to support engineers and stakeholders to make a well-informed decision when procuring LLM infrastructure.

The performance of the LLM is a combination of model characteristics, hardware capabilities, and software efficiency.

Use cases

The following use cases are considered, based on the LocalScore benchmark:

Use case	Description	Prompt Tokens	Text Generation Tokens
UC1	Classification, sentiment analysis, keyword extraction	1024	16
UC2	Long document Q&A, RAG, short summary of extensive text	4096	256
UC3	Complex reasoning, chain-of-thought, long-form creative writing, code generation	1280	3072
UC4	Prompt expansion, explanation generation, creative writing, code generation	384	1152

Levels of interactiveness

A workflow is considered -

- *Interactive workflow*: Input tokens are processed in less than 5 seconds, and the output is produced at more than 10 tokens/s.
- *Background task*: Input processing + output generation finishes within 300 seconds (5 minutes).
- *Off-line workflow*: Input processing + output generation finishes within a few hours / overnight. In this case a progress indicator is helpful.
- *Non-viable*, if the task takes prohibitively long time to finish.

Hardware platforms:

We examine four types of AMD platforms: server, workstation, desktop, and embedded/mobile.

Platform	AMD CPU example	Cores	RAM channels	RAM [GB]	RAM WB [GB/s]	Max GPUs	Max GPU VRAM [GB]	Max ML model size [B params]
Server	EPYC 9554P	64	12	384	460.8	6 (8)	576	300
Workstation	Threadripper 7970X	32	4	256	166.4	2 (6)	192	80
Desktop	Ryzen 9 9950X	16	2	192	89.6	1 (2)	96	40
Embedded	Ryzen AI MAX+ PRO 395	16	2	128	128	integrated	96	40

The following hardware configurations are examined:

AMD CPU example	GPU type	GPU VRAM [GB]	ML model size [GB]	Prompt processing [token/s]	Token generation [token/s]	UC1	UC2	UC3	UC4
EPYC 9554P	-	-	10	223?	21?	√	-	√	?
EPYC 9554P	4 x RTX PRO 6000 Blackwell	384	max 192	?	?	√	√	√	√
Threadripper 7970X	-	-	10	223	14	√	-	√	?
Threadripper 7970X	RTX PRO 6000 Blackwell	96	10	5126	81	√	√	√	√
Threadripper 7970X	2 x RTX PRO 6000 Blackwell	192	max 96	?	?	√	√	√	√
Ryzen 9 9950X	-	-	10	125	8	√	-	?	?
Ryzen 9 9950X	RTX 5080	16	10	2291	24	√	√	?	?
Ryzen 9 9950X	RTX 5090	32	10	4787	65	√	√	√	√
Ryzen AI MAX+ PRO 395	-	-	10	84	11	√	-	-	?

Characteristics of running large language models

The large language model runner takes a sequence of tokens as input (called a prompt) and produces a sequence of tokens as output. Running an LLM is also called *inferencing*.

Important parameters are:

- Model size
  - Larger models generally achieve higher accuracy and better performance, but require more memory and computational resources.
- Model quantization
  - Lower bit depth (e.g., 8-bit, 4-bit) reduces memory and computation needs but can decrease inference accuracy. Higher bit quantization (16-bit, 32-bit) preserves precision but requires more memory and processing.

- Model architecture
- Prompt length
    - Longer prompts require more computation
- Prompt batch size
    - Processing multiple prompts in parallel can improve throughput if the CPU has enough computing power.

Model Size

Larger models generally achieve higher accuracy and better performance, but require more memory and computational resources. Smaller models are faster, and use less memory, but may have lower accuracy and are worse at complex tasks. The trade-off is between performance and resource usage.

The size of an LLM is typically measured by the number of parameters -- the weights and biases in the artificial neural network. LLMs typically have millions or billions of parameters, and the total storage size is measured in gigabytes.

S_{model}: Size of the model file in GB. Typical values for publicly available language models:

Model Name	Parameters	S_{model} FP16
Tiny models (e.g. DistilGPT2)	~80M	~0.15 GB
GPT-2 Small	124M	~0.25 GB
GPT-2 Medium	345M	~0.70 GB
GPT-2 Large	774M	~1.5 GB
GPT-2 XL	1.5B	~3.0 GB
Llama 2 7B	7B	~13 GB
Llama 2 13B	13B	~25 GB
Llama 2 70B	70B	~140 GB
Falcon 180B	180B	~360 GB

For number representation FP16 (16-bit floating point) is commonly used for inference, halving the storage compared to FP32, typically used when training the model. Actual file sizes vary slightly due to model architecture and tokenizer data. Some models (e.g., Mistral, Phi-2, Gemma) fall between these sizes (e.g., 2–8 GB for 1–3B parameters).

Model quantization

Lower bit depth (e.g., 8-bit, 4-bit) reduces memory and computation needs but can decrease inference accuracy. Higher bit quantization (16-bit, 32-bit) preserves precision but uses more resources.

Typical quantization values used in local LLMs are 4-bit (int4), 8-bit (int8), and sometimes 16-bit (fp16 or bfloat16.) The 4-bit and 8-bit quantizations are most common for efficient inference, while 16-bit is often used for training or higher-precision inference.

There is no universal optimum bit depth -- it depends on the model, task, and expected accuracy. 32-bit models (fp32) are used when maximum precision is needed, such as during model development, debugging, or when training very large models where numerical stability is critical. 16-bit models (fp16 or bfloat16) are commonly used during training to save memory and speed up computation while maintaining good accuracy. For inference, 16-bit is often sufficient, but 32-bit may be used if highest accuracy is required. 8-bit is often a good balance for inference, while 4-bit is used for maximum efficiency with some accuracy trade-off.

Prompt length

L_{prompt}: Number of tokens in the input prompt.

Longer prompts require more computation, especially for attention mechanisms.

- Short prompts: 5–50 tokens (e.g., simple questions, instructions)
- Medium prompts: 50–200 tokens (e.g., paragraphs, multi-step instructions)
- Long prompts: 200–1000+ tokens (e.g., documents, code, multi-turn conversations)

Most practical prompts are between 10 and 200 tokens. Some advanced applications (retrieval-augmented generation, long context tasks) may use much longer prompts, up to the model's context window (often 2,048–32,768 tokens, or even more, depending on the model).

Prompt processing time grows with the second power of prompt length. Token generation time grows linearly with prompt length, offset by the model size.

Prompt batch size

B: Batch size - number of prompts processed in parallel. Batching can improve throughput if the CPU has enough computing power. Typical batch sizes for LLM inference depend on the use case and available system resources:

- For interactive/chat use, $B = 1$
  - This is the most common if low latency is prioritized
- For batch processing / offline inference: $B = 2..16$ (sometimes up to 32 or 64 on high-end CPUs/GPUs)
- For CPU-based LLMs: Batch sizes are usually $B = 1..4$ due to memory and bandwidth constraints
- For GPU-based LLMs: Batch sizes can be much larger ($8..128+$), limited by GPU memory

Required operations per output token

OP_{token}: Number of floating point operations needed to process/generate one token [FLOP/token]. As a rule of thumb:

$$ OP_{token} [FLOP/token] \approx 2 \times N \times d_{model}^2 $$

Where:

- N: Number of transformer layers, and \
- d_{model}: hidden size.

Typical Values:

Model	Op per token [GFLOP/token]
GPT-2 Small (124M)	2–3
GPT-2 Medium (345M)	6–8
GPT-2 Large (774M)	15–20
GPT-3 6.7B	120–150
GPT-3 13B	250–300
Llama 7B	80–100
Llama 13B	160–200
Llama 70B	800–1000

These are rough estimates; actual values depend on architecture details and prompt length (due to attention scaling).

LLM inference performance indicators

The following parameters are indicators of inference performance:

- Prompt processing throughput [token/second]
  - Number of input tokens processed in a second.
- Time to first token [millisecond] (depends on prompt processing throughput plus input prompt length)
  - The elapsed time between submitting the prompt and receiving the first output token.
- Token generation throughput [token/second]
  - Number of tokens generated in a second after emitting the first token.

Prompt Processing Throughput

PP: Number of input tokens processed in a second. The model attends to all prompt tokens (full attention over the entire prompt). Computational cost grows with the second power of prompt length ($O(L_{prompt}^2)$ for attention).

The formula for prompt processing speed (token/s) is:

$$ PP [token/s] = \dfrac{ F_{cpu} \times E_{par} }{ OP_{token} } $$

Where:

- F_{cpu}: CPU FLOPS (floating-point operations per second) $[FLOP/s]$
- E_{par}: Parallelism efficiency (0–1) $[1]$
- OP_{token}: Floating point operations required to process one token $[FLOP/token]$

This formula assumes the process is compute-bound. For very large models, memory bandwidth also becomes a limiting factor.

Time to First Token

T_{first}: The time from submitting a prompt to receiving the first output token. It includes the time to process the input prompt (also called *prefill*), and the time needed to generate the first output token.

A simplified formula:

$$ T_{first} [s] \approx \dfrac{L_{prompt}}{PP} + \dfrac{S_{model}}{BW_{mem}} $$

Where:

- L_{prompt}: Prompt length (number of tokens) $[token]$
- The first term is the compute time for the prompt (prefill)
- The second term is the time to load model weights from RAM (shorter if already cached)

Token Generation Throughput

TG: Token Generation Throughput. Number of tokens generated in a second after emitting the first token. For large models, token generation is often memory-bound:

$$ TG [token/s] \approx \dfrac{BW_{mem} [GB/s]}{S_{model} [GB]} $$

For example, if memory bandwidth is 50 GB/s and model size is 10 GB: $TG \approx 50 / 10 = 5~token/s$. If compute is the bottleneck (for small models or slow CPUs):

$$ TG [token/s] \approx \dfrac{(F_{cpu} × E_{par})}{OP_{token}} $$

For most large LLMs on CPUs, memory bandwidth and model size are the main determinants for token generation throughput. Prompt processing (time to first token) is more compute-bound and depends on prompt length and processing speed.

Local LLM inference hardware

LLM inference prefill (preprocessing) is compute-bound: The calculation speed matters the most: processor clock frequency, and cache size (L3 and L1). Token generation is memory bandwidth-bound: Higher memory throughput results in faster token generation.

GPUs have both high computing power and high memory bandwidth. A GPU with even a small amount of memory can accelerate inference speeds considerably.

The system RAM should be large enough to accommodate the model plus the context. If the system RAM is too small, then swapping to (relatively slow) SSD will degrade overall performance.

The following hardware configurations are considered:

- CPU only
- CPU + GPU: The model plus context fits into the GPU VRAM
- Hybrid CPU + GPU: The model plus context fits into the CPU memory, but does not fit into the GPU VRAM. Typical examples are mobile solutions (laptop, mini PC).

Processing Performance

F_{cpu}: CPU floating-point operations per second (FLOPS or FLOP/s). How quickly the computer can process data is directly affected by the number of cores, clock speed, cache size, and architecture (notably, AVX2/AVX-512 support).

- Modern Desktop CPUs (2023–2025):
  - Single Core: $~20–100$ GFLOPS (billion FLOPS)
  - Multi-core ($`8..32`$ cores): $~200..2000$ GFLOPS ($`0.2`$ to $`2`$ TFLOPS)
- High-End Server CPUs:
  - Multi-socket, AVX-512: Up to $`5..10`$ TFLOPS (theoretical peak, rarely sustained in practice)

Actual sustained FLOPS for LLM inference is usually lower due to memory bottlenecks and non-ideal vectorization.

CPU performance calculation example

The Intel Core i5-1335U is a 13th Gen mobile CPU with a hybrid architecture (Performance and Efficiency cores). Intel does not publish official FLOPS (floating point operations per second) numbers for consumer CPUs, but one can estimate peak theoretical FLOPS per core as follows:

1. Max clock speed per core:

   - P-cores: up to 4.6 GHz
   - E-cores: up to 3.4 GHz

2. SIMD width:

   - Supports AVX2 (256-bit SIMD), which is 8 single-precision (FP32) or 4 double-precision (FP64) operations per instruction.

3. FLOPS per core (theoretical, FP32):

   - Each core can do 8 FP32 operations per cycle (with AVX2 FMA, fused multiply–add operations in one clock cycle).
   - FLOPS = SIMD width × FMA × clock speed

Core Type	Max Clock	Theoretical Peak FP32 FLOPS per Core
P-core	4.6 GHz	~73.6 GFLOPS
E-core	3.4 GHz	~54.4 GFLOPS

Calculation:

- P-core at 4.6 GHz: $FLOPS = 8 (SIMD ~ FP32 ops) × 2 (FMA) × 4.6e9 (Hz) = 73.6~FP32~GFLOPS$
- E-core at 3.4 GHz: $FLOPS = 8 × 2 × 3.4e9 = 54.4~FP32~GFLOPS$
- Total: $2 \times 73.6 + 8 \times 54.4 = 582.4~FP32~GFLOPS$
- The actual measured value is 42.2 GFLOPS, 7% of the theoretical maximum.

These are theoretical maximums assuming AVX2 FMA is fully utilized, which is rare in real-world workloads. Actual FLOPS will be lower due to memory, instruction mix, and other bottlenecks.

Disable hyperthreading: Use only one thread per CPU core

For LLM inference on CPUs, it is best to use one thread per physical core.

Hyperthreading (SMT) usually provides limited or no benefit, and sometimes even reduces performance compared to running one thread per physical core. This is because LLM inference is typically memory bandwidth-bound. Hyperthreading helps when there are idle CPU resources (e.g., waiting on memory), but with LLMs, all threads often compete for the same memory bandwidth. Adding more threads can increase contention and cache thrashing.

Thread/Parallelism Efficiency

E_{par}: How well the workload is parallelized. It is the ratio of actual speedup to theoretical maximum speedup. Typical values on a multi-core CPU are:

- Ideal (perfect scaling): $E_{par} = 1.0$
- Real-world (good scaling): $E_{par} = 0.7$ to 0.9
- Suboptimal (memory-bound or poor threading): $E_{par} = 0.4$ to 0.7

For example, if the CPU has 16 cores and the observed speedup is 10x over a single core, then $E_{par} = 10/16 = 0.625$. For LLM inference, E_{par} is often 0.6 to 0.85 on modern CPUs, depending on model size, memory bandwidth, and software optimization. Larger models and memory-bound workloads tend to have lower values.

CPU-only inference

For LLM inference the CPU needs to support the AVX-512 instruction set's bfloat16 format. In the case of AMD CPUs, it is available in the Zen~4 and Zen~5 architectures.

Maximum theoretical system memory bandwidth calculation

The maximum theoretical system memory bandwidth is determined by the memory modules' speed, how many memory channels there are, and how fast the CPU can exchange data with the memory modules.

$$ BW_{RAM} = RAM channels * 8 * RAM speed [GT/s] $$

$$ BW_{CPU} = n_{CCD} per core * 32 * FCLK [GHz] $$

$$ BW = min(BW_{RAM}, BW_{CPU}) $$

Where:

- BW_{RAM}: Theoretical bandwidth on the RAM modules' side
- BW_{CPU}: Theoretical bandwidth on the CPU memory controller side
- BW: Theoretical system RAM bandwidth.
    - Actual values are measured between 30% to 95% of the theoretical maximum.
- FCLK: Fabric Clock Speed - clock speed of the memory controller in the CPU.
    - Zen 4: 1.8 Ghz, Zen 5: 2.0 GHz. Some models can be overclocked.
- CCD: Core complex Die - contains the memory controller on the CPU side.
    - Ranges from 1 to 16 per CPU core.
- RAM speed: 4.8 GT/s to 8 GT/s; can be overclocked
- RAM channels: 2-12 (24 for 2-CPU setups).
    - RAM sizes vary between 4-128 GB
    - Viable total memory size: 8 GB to 1536 GB. (3072 GB in the case of 24 x 128 GB modules.)
- See also fairydreaming's recommendation on Reddit:

> To get the best memory bandwidth, (theoretically) you should:
>
> - Increase FCLK for 8-channel configurations with 2 or 4 CCDs (7945WX, 7955WX, 7965WX, 7975WX),
> - Use overclocked memory in all remaining Threadripper models,
> - For Epyc, purchase a motherboard with 12 memory slots and an Epyc 9004 processor with at least 8 CCDs. Fill all memory slots.

Examples:

- AMD Ryzen 5 7400F: n_CCD = 1, FCLK = 1.8 GHz, ch = 2.
    - BW = min(2 ch * 8 * 5.2 GT/s, 1 CCD * 32 * 1.8 GHz) = min(83.2, 57.6) GB/s = 57.6 GB/s
    - If the model size plus context is 10 GB, then the generation throughput is less than 5.8 token/s.
- AMD EPYC 9755: n_CCD = 16, FCLK = 2.0 GHz, ch = 12.
    - BW = min(12 ch * 8 * 5.6 GT/s, 16 CCD * 32 * 2.0 GHz) = min(537.6, 1024) GB/s = 537.6 GB/s
    - If the LLM size plus context is 10 GB, then the generation throughput is less than 53.8 token/s.

System memory bandwidth links:

- https://www.reddit.com/r/LocalLLaMA/comments/1fcy8x6/memory_bandwidth_values_stream_triad_benchmark/
- https://www.reddit.com/r/LocalLLaMA/comments/1h3doy8/stream_triad_memory_bandwidth_benchmark_values/

GPU parameters

- Memory (VRAM): Determines the maximum model size one can run. Larger models and larger contexts require more VRAM.

- Memory bandwidth is very important for feeding data to compute units.

- Compute Units/Shading Units: They perform the matrix operations. More units generally mean faster inference (The compute units are called CUDA cores on NVIDIA, Stream Processors on AMD.)

- Tensor Cores/Matrix Cores: Specialized hardware that dramatically accelerates matrix operations

- Data type support: Hardware support for FP16, BF16, INT8 operations

GPU parameters not relevant for LLM inference:

- TMUs (Texture Mapping Units): Graphics-specific feature for texture handling
- ROPs (Render Output Pipelines): Handle pixel output operations for displays
- RT Cores (Ray Tracing Cores): Specialized for graphics ray tracing

GPU Links:

- https://www.reddit.com/r/LocalLLaMA/comments/1k57b1o/comment/moftfs0/?context=3

Memory Bandwidth and Latency

BW_{mem}: RAM bandwidth in GB/s. LLMs are memory-intensive; insufficient bandwidth limits performance.

First, the LLM model file needs to be loaded from the disk into the system memory. This is a one-off initialization step, bound by the disk's throughput. The typical read bandwidth of a modern SSD ranges from 500 MB/s (SATA SSDs) to 3,000–7,000 MB/s (NVMe SSDs using PCIe Gen3/Gen4). In 2025, the fastest SSDs can achieve a peak throughput of 14 GB/s. Assuming 1-10 GB/s SSD read speed and 1-100 GB LLM model files sizes, loading the model to the system RAM can range from 0.1 second to 1.5 minutes -- typically a few seconds.

If CPU-only inference is used, then each weight and bias of the model is loaded to the CPU's L3, L2, and L1 cache stage-by-stage. Depending on how well the inference code is optimized, this loading procedure may need to be repeated for each output token produced. In turn, the system's inference speed is determined by how many times per second the complete model plus the context can be loaded into the memory.

In the case of GPU inference the process is similar: the model is first loaded from the system RAM into the GPU VRAM; then subsequently into the GPU Compute Units' L3, L2, and L1 cache. As long as the GPU has sufficient VRAM to hold the complete model, the token generation speed is determined by the GPU VRAM throughput.

The inference process requires matrix multiplications, which are relatively cheap operations compared to the time required to move data within memory. For details, see Tim Dettmers' blog post Which GPU(s) to Get for Deep Learning: My Experience and Advice for Using GPUs in Deep Learning.

If the model does not fit completely into the GPU VRAM, then hybrid CPU+GPU operation is needed. System performance is determined by how much data is moved about within system RAM.

Typical memory throughput values:

- CPU only:

    - DDR4: 25–50 GB/s (dual/quad channel)
    - DDR5: 50–100 GB/s (higher end, e.g., Intel 13th/14th Gen, AMD Ryzen 7000+)
- New APUs (AMD Ryzen 8000G, AMD Ryzen AI Max+ PRO 395, Intel Meteor Lake): - Use fast LPDDR5/DDR5, typically 50–200 GB/s shared between CPU and integrated GPU
- GPU only:
    - Consumer GPUs (e.g., RTX 4060): 250–400 GB/s
    - High-end GPUs (e.g., RTX 4090, A100): 800–2000+ GB/s (with GDDR6X or HBM2/3)
- CPU+GPU:
    - CPU and GPU bandwidths are added separately; data transfer between them is limited by PCIe (16–64 GB/s for PCIe 4.0/5.0 x16)

Real-world bandwidth is often lower than peak theoretical values -- observed bandwidth can be as low as 0.3 to 0.8 of the theoretical maximum. For AI/LLM workloads, GPU bandwidth is usually the bottleneck. (See STREAM TRIAD benchmark results.)

In the next sections the following CPU categories are detailed:

- Server CPUs
- Workstation CPUs
- Desktop CPUs
- Embedded CPUs

Server CPUs

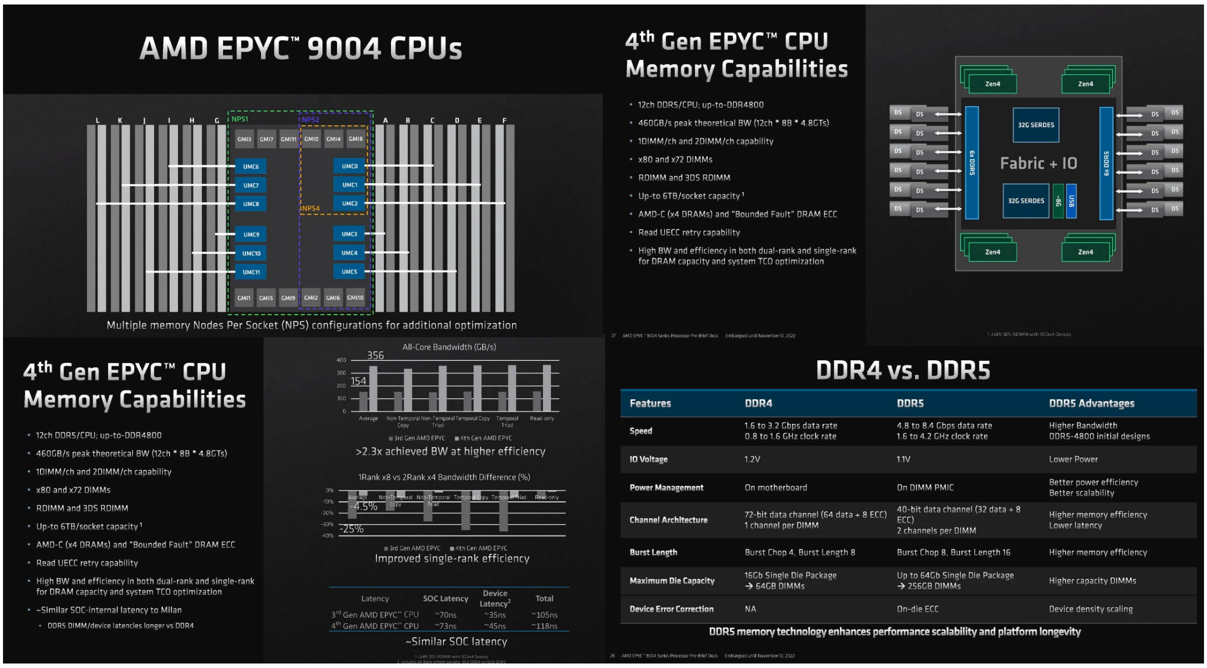
- Socket: SP5:
    - EPYC 9004: 12-channel DDR5-4800 (Zen 4) (24 for 2-CPU config)
    - EPYC 9005: 12-channel DDR5-5600 (Zen 5) (24 for 2-CPU config)
- Maximum RAM: 3 TB (6 TB for 2-CPU config)
- Maximum cores: 192 (counts at prefill throughput)
- Cache:
    - L1: 80 KB (48 KB data + 32 KB instruction) per core.
    - L2: 1 MB per core
    - L3: 16-1152 MB per CPU
- Maximum PCIe lanes: 128 (160 in 2-CPU config)
    - Enough to add 8 CPIe 5.0 x16 GPUs (TechPowerUp)
- Maximum number of CCDs per CPU: 16
    - Minimum 9 CCDs are required to serve the RAM module bandwidth
        - Assert $9~CCD \times 32 \times 2.0~GHz \ge 12~ch \times 8 \times 5.6~GT/s$
            - See calculation
- Maximum theoretical memory bandwidth: 460.8 to 1075.2 GB/s

    - Caps token generation throughput of 10 GB LLM model at 46 to 100 token/s

    - Note that from the Genoa platform on, single-rank memory modules will perform well

      > The other important feature is dual rank versus single rank memory. With Milan and most Intel platforms, dual-rank memory is crucial to maximizing performance. There's a 25% performance delta on Milan, for example. With Genoa, this is brought down to 4.5%. This is another considerable cost improvement because cheaper single-rank memory can be used.

      Slide:



AMD EPYC™ 9004 CPUs

Multiple memory Nodes Per Socket (NPS) configurations for additional optimization

4th Gen EPYC™ CPU Memory Capabilities

- 12ch DDR5/CPU; up-to-DDR4800
- 460GB/s peak theoretical BW (12ch * 8B * 4.8GTs)
- 1DIMM/ch and 2DIMM/ch capability
- x80 and x72 DIMMs
- RDIMM and 3DS RDIMM
- Up-to 6TB/socket capacity [1]
- AMD-C (x4 DRAMs) and "Bounded Fault" DRAM ECC
- Read UECC retry capability
- High BW and efficiency in both dual-rank and single-rank for DRAM capacity and system TCO optimization

27 AMD EPYC™ 9004 Series Processor Pre-Brief Deck Embargoed until November 10, 2022

[1] 2x8R 3DS-RDIMM with 16Gbx4 Devices

4th Gen EPYC™ CPU Memory Capabilities

- 12ch DDR5/CPU; up-to-DDR4800
- 460GB/s peak theoretical BW (12ch * 8B * 4.8GTs)
- 1DIMM/ch and 2DIMM/ch capability
- x80 and x72 DIMMs
- RDIMM and 3DS RDIMM
- Up-to 6TB/socket capacity [1]
- AMD-C (x4 DRAMs) and "Bounded Fault" DRAM ECC
- Read UECC retry capability
- High BW and efficiency in both dual-rank and single-rank for DRAM capacity and system TCO optimization
- ~Similar SOC-internal latency to Milan
    - DDR5 DIMM/device latencies longer vs DDR4

>2.3x achieved BW at higher efficiency

Improved single-rank efficiency

Latency	SOC Latency	Device Latency[2]	Total
3rd Gen AMD EPYC™ CPU	~70ns	~35ns	~105ns
4th Gen AMD EPYC™ CPU	~73ns	~45ns	~118ns

~Similar SOC latency

[1] 2x8R 3DS-RDIMM with 16Gbx4 Devices
[2] Includes All Bank refresh penalty, 8Gb DDR4 vs 16Gb DDR5

DDR4 vs. DDR5

Features	DDR4	DDR5	DDR5 Advantages
Speed	1.6 to 3.2 Gbps data rate 0.8 to 1.6 GHz clock rate	4.8 to 8.4 Gbps data rate 1.6 to 4.2 GHz clock rate	Higher Bandwidth DDR5-4800 initial designs
IO Voltage	1.2V	1.1V	Lower Power
Power Management	On motherboard	On DIMM PMIC	Better power efficiency Better scalability
Channel Architecture	72-bit data channel (64 data + 8 ECC) 1 channel per DIMM	40-bit data channel (32 data + 8 ECC) 2 channels per DIMM	Higher memory efficiency Lower latency
Burst Length	Burst Chop 4, Burst Length 8	Burst Chop 8, Burst Length 16	Higher memory efficiency
Maximum Die Capacity	16Gb Single Die Package → 64GB DIMMs	Up to 64Gb Single Die Package → 256GB DIMMs	Higher capacity DIMMs
Device Error Correction	NA	On-die ECC	Device density scaling

DDR5 memory technology enhances performance scalability and platform longevity

26 AMD EPYC™ 9004 Series Processor Pre-Brief Deck Embargoed until November 10, 2022

Series (SP5 socket)	Cores	Max RAM size (1/2-CPU)	Max RAM BW (1/2-CPU)	Token generation (1/2-CPU)
EPYC 9004	16-128	3 / 6 TB	460.8 / 921.6 GB/s	15-46 / 21-72 token/s

Series (SP5 socket)	Cores	Max RAM size (1/2-CPU)	Max RAM BW (1/2-CPU)	Token generation (1/2-CPU)
EPYC 9005	6-192	3 / 6 TB	537.6 / 1075.2 GB/s	18-53 / 25-86 token/s

Test results at OpenBenchmarking.org for llama.cpp and LocalScore.

Prices for a complete computer with SP5 CPU socket are in the 2,500--7,000 EUR range.

Server CPU links:

- https://www.amd.com/en/products/specifications/processors.html
- https://www.amd.com/en/products/specifications/server-processor.html
- https://en.wikipedia.org/wiki/Epyc#Fifth_generation_Epyc_(Turin_and_Turin_Dense)
- https://forum.level1techs.com/t/amd-epyc-9005-series-sp5-single-slot-or-dual-slot-motherboard-suggestion/229954
- https://www.techpowerup.com/cpu-specs/
- https://www.techpowerup.com/327388/amd-granite-ridge-zen-5-processor-annotated
- https://www.thomas-krenn.com/en/wiki/Optimize_memory_performance_of_Intel_Xeon_Scalable_systems
- https://www.servethehome.com/memory-bandwidth-per-core-and-per-socket-for-intel-xeon-and-amd-epyc/

Workstation CPUs

This includes the AMD Ryzen Threadripper and EPYC 8004 processors.

- Sockets:
  - sTR5:
    - Threadripper: 4-channel DDR5-5200/6400 (Zen 4/5)
    - Threadripper Pro: 8-channel DDR5-5200/6400 (Zen 4/5)
  - SP6:
    - EPYC 8004: 6-channel DDR5-4800 (Zen 4 only)
- Maximum RAM: 1024 GB (2048 GB?)
- Maximum cores: 96 (counts at prefill throughput)
- Cache:
  - L1: 80 KB (48 KB data + 32 KB instruction) per core.
  - L2: 1 MB per core
  - L3: 64-384 MB per CPU
- Maximum PCIe lanes:
  - Threadripper: 48 PCIe 5.0 and 24 PCIe 4.0
    - Enough to add 2 CPIe 5.0 x16 GPUs (TechPowerUp)
  - Threadripper Pro: 128 PCIe 5.0 lanes
    - Enough to add 6 CPIe 5.0 x16 GPUs (TechPowerUp)
- Maximum number of CCDs per CPU: 12
- Maximum theoretical memory bandwidth: 166.4 to 409.6 GB/s
  - Caps token generation throughput of 10 GB LLM model plus context at 16 to 40 token/s

CPU Series	Socket	Cores	Max RAM size	Max RAM BW	Token generation
Threadripper 7000X	sTR5	24-64	512 GB	166.4 GB/s	5-16 token/s
Threadripper Pro 7000WX	sTR5	12-96	1024 GB	332.8 GB/s	10-33 token/s
Threadripper 9000X	sTR5	24-64	512 GB	204.8 GB/s	6-20 token/s
Threadripper Pro 9000WX	sTR5	12-96	2048 GB	409.6 GB/s	13-40 token/s
EPYC 8004	SP6	8-64	768 GB	230.4 GB/s	7-23 token/s

Prices for a complete workstation computer are in the 10,000--20,000 EUR range.

Workstation CPU links:

- https://www.amd.com/en/products/processors/workstations/ryzen-threadripper.html
- https://en.wikipedia.org/wiki/Threadripper#Shimada_Peak_(Threadripper_9000_series,_Zen_5_based)
- https://www.techpowerup.com/cpu-specs/epyc-8024pn.c3295

Desktop CPUs

- Socket: AM5 (dual-channel RAM)
- Maximum RAM: 256 GB (See Level1Tech1: 256 GB on AM5 on YouTube)
- Maximum cores: 16 (determines prefill throughput)
- Maximum L3 cache: 32-128 MB (limits prefill throughput)
  - L1 cache (Zen 4/4c): 32+32 kB, L2: 1 MB, L3: 32 MB per CCD.
  - L1 cache (Zen 5/5c): 32+48 kB, L2: 1 MB, L3: 32 MB per CCD.
- Maximum available PCIe lanes: 24.
  - Enough to handle only one CPIe 5.0 x16 GPU
- Maximum theoretical memory bandwidth: 89.6 GB/s
  - This caps token generation for a 10 GB LLM model + context at 9 token/s.

CPU Series (AM5 socket)	Cores	Max RAM size	Max RAM BW	Token generation
Ryzen 7700/7900/7950	6-16	128 GB	83.2 GB/s	3-8 token/s
Ryzen 7040/7045/8000F/8000G	6-16	128 GB	83.2 GB/s	3-8 token/s
EPYC 4004	4-16	128 GB	83.2 GB/s	3-8 token/s
Ryzen 9000	6-16	192 GB	89.6 GB/s	3-8 token/s

Prices for a complete desktop computer are in the 1000--5000 EUR range.

Desktop CPU links:

- https://www.anandtech.com/show/21524/the-amd-ryzen-9-9950x-and-ryzen-9-9900x-review/10
- https://www.amd.com/en/products/processors/desktops/ryzen/9000-series/amd-ryzen-9-9950x3d.html
- https://www.amd.com/en/products/processors/chipsets/am5.html
- https://skatterbencher.com/2025/03/11/skatterbencher-85-ryzen-9-9950x3d-overclocked-to-5900-mhz/
- Level1Tech1: 256 GB on AM5

  > **Case**: Fractal North (glass or mesh side panel options)
  >
  > **Motherboard**: MSI X870 Tomahawk (chosen for best 256GB RAM compatibility)
  >
  > **CPU**: AMD 9950X3D (16-core)
  >
  > **Memory**: 256GB G.Skill Trident Z Neo DDR5-6000 (4x64GB)
  >
  > **CPU Cooler**: Noctua NH-D15 (with FSP MP7 suggested as alternative)
  >
  > **GPU**: AMD 7900 XTX (24GB VRAM)
  >
  > **Power Supply**: 850W (CIC or FSP HydroG recommended)
  >
  > **Storage**: Crucial T710 NVMe recommended, also MemBlaze PBlaze7 7A40 NVMe SSD

Mobile CPUs with integrated NPU

- Sockets: FL1, FP7, FP7r2 or FP8 type packages
  - 200: All models support DDR5-5600 or LPDDR5X-7500 in 128-bit "dual-channel" mode.
  - 300: All models support DDR5-5600 or LPDDR5X-8000 in dual-channel mode.
- Maximum RAM: 128 GB
- Maximum cores: 4-12 (counts at prefill throughput)
- Cache:
  - L1: 80 KB (48 KB data + 32 KB instruction) per core.
  - L2: 1 MB per core
  - L3: 8-64 MB
- Maximum PCIe lanes: 16-20
- Maximum theoretical memory bandwidth: 128 GB/s
  - This caps token generation throughput of a 10 GB LLM model at 12 token/s

CPU Series (embedded)	Cores type	Max RAM	Max RAM BW	Token generation
Ryzen 8040	4-8	128 GB	89.6 GB/s	3-9 token/s
Ryzen AI 200	4-8	256 GB	128 GB/s	4-12 token/s
Ryzen AI 300	4-12	256 GB	128 GB/s	4-12 token/s
Ryzen AI MAX/MAX+	6-16	128 GB	128 GB/s	4-12 token/s

Prices for a complete computer are in the 1,000--2,500 EUR range.

Embedded CPU Links:

- https://en.wikipedia.org/wiki/List_of_AMD_Ryzen_processors#Ryzen_AI_300_series
- https://www.techpowerup.com/cpu-specs/ryzen-ai-max-pro-395.c3998
- https://www.amd.com/en/products/processors/laptop/ryzen-pro/ai-max-pro-300-series/amd-ryzen-ai-max-plus-pro-395.html
- https://www.amd.com/en/blogs/2025/amd-ryzen-ai-max-395-processor-breakthrough-ai-.html
- https://www.tomshardware.com/pc-components/cpus/more-affordable-strix-halo-model-emerges-early-ryzen-ai-max-385-geekbench-result-reveals-an-eight-core-option

CPU cooler

CPU cooler links:

- https://noctua.at/en/products/fan/nf-a12x25-ls-pwm
- https://www.thermaltake.com/aw420-aio-liquid-cooler.html
- https://thermaltakeusa.com/products/aw360-aio-liquid-cooler-cl-w450-pl12bl-a
- https://www.silverstonetek.com/en/product/info/coolers/xed120s_ws/
- https://www.silverstonetek.com/en/product/info/coolers/xe04_sp5/
- https://forums.servethehome.com/index.php?threads/cooler-recommendations-for-400w-sp5.43530/
- https://noctua.at/en/nh-d9-tr5-sp6-4u/specification
- https://www.coolserver.com.cn/en/product_view_397_283.html
- https://www.arctic.de/en/Freezer-4U-SP5/ACFRE00158A
- https://www.phoronix.com/review/arctic-freezer-4u-sp5
- https://www.coolserver.com.cn/en/product_view_598_283.html

Benchmarks

Benchmarking and aggregator sites

- OpenBenchmarking.org: Storage of Phoronix Test Suite benchmark result data
  - Machine Learning Test Suite: Popular pattern recognition and computational learning algorithm benchmarks
    - AI Benchmark Alpha, Caffe, LeelaChessZero, LiteRT, Llama.cpp, Llamafile, LocalScore, Mlpack Benchmark, Mobile Neural Network, NCNN, Neural Magic DeepSparse, Numenta Anomaly Benchmark, Numpy Benchmark, oneDNN, ONNX Runtime, OpenCV, OpenVINO, OpenVINO GenAI, PlaidML, PyTorch, R Benchmark, Scikit-Learn, spaCy, TensorFlow, TensorFlow Lite, Whisper.cpp
- LocalScore: Benchmarks for LLMs and a repository for the results.

  > A LocalScore is a measure of three key performance metrics that matter for local LLM performance: Prompt Processing Speed, Generation Speed, and Time to First Token. These metrics are combined into a single LocalScore which gives you a straightforward way to compare different hardware configurations. A score of 1,000 is excellent, 250 is passable, and below 100 will likely be a poor user experience in some regard.

  - Models:

Model Size	Tiny	Small	Medium
# Params	1B	8B	14B
Model	Llama 3.2	Llama 3.1	Qwen 2.5
Quantization	Q4_K_M	Q4_K_M	Q4_K_M
Approx VRAM Required	2 GB	6 GB	10 GB

  - Use cases:

Prompt Tokens	Text Generation Tokens	Sample Use Cases
1024	16	Classification, sentiment analysis, keyword extraction
4096	256	Long document Q&A, RAG, short summary of extensive text
2048	256	Article summarization, contextual paragraph generation

Prompt Tokens	Text Generation Tokens	Sample Use Cases
2048	768	Drafting detailed replies, multi-paragraph generation, content sections
1024	1024	Balanced Q&A, content drafting, code generation based on long sample
1280	3072	Complex reasoning, chain-of-thought, long-form creative writing, code generation
384	1152	Prompt expansion, explanation generation, creative writing, code generation
64	1024	Short prompt creative generation (poetry/story), Q&A, code generation
16	1536	Creative text writing/storytelling, Q&A, code generation

- Discussions:
  - https://www.reddit.com/r/LocalLLaMA/comments/1iyztni/dual_9175f_amd_epyc_9005_a_new_trend/
  - https://www.reddit.com/r/LocalLLaMA/comments/1jq13ik/comment/ml6hg70/?context=3
  - https://www.reddit.com/r/threadripper/comments/1azmkvg/comparing_threadripper_7000_memory_bandwidth_for/
  - https://old.chipsandcheese.com/2024/11/24/pushing-amds-infinity-fabric-to-its-limits/
  - https://www.servethehome.com/amd-epyc-genoa-gaps-intel-xeon-in-stunning-fashion/3/
  - https://www.reddit.com/r/FlowZ13/comments/1j2uymr/comment/mhauazc/
  - https://www.reddit.com/r/LocalLLaMA/comments/1kedbv7/ryzen_ai_max_395_a_gpu/
  - https://www.reddit.com/r/LocalLLaMA/comments/1kmi3ra/amd_strix_halo_ryzen_ai_max_395_gpu_llm/
  - https://www.reddit.com/r/LocalLLaMA/comments/1ghvwsj/llamacpp_compute_and_memory_bandwidth_efficiency/
  - https://www.reddit.com/r/LocalLLaMA/comments/1ghvwsj/comment/lv4sx1e/

- LocalScore results on OpenBenchmarking.org
  - Models:

FileName	FileSize
localscore-0.9.3	380,479,144
Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf	4,920,739,232
Qwen2.5-14B-Instruct-Q4_K_M.gguf	8,988,110,976
Qwen3-32B-Q4_K_M.gguf	19,762,150,048

  - A few selected results:
    - Model: Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf - Acceleration: CPU

Component	Prompt Processing [token/s]	Time To First Token [ms]	Token Generation [token/s]
AMD Ryzen 9 9950X 16-Core	246	5660 +/- 2	14.5
AMD Ryzen 9 9950X3D 16-Core	236 +/- 2	5831 +/- 46	14.3 +/- 0.1
AMD EPYC 4585PX 16-Core	226 +/- 30		11.3 +/- 0.1
AMD EPYC 4465P 12-Core	165	8415 +/- 5	11.9
AMD RYZEN AI MAX+ PRO 395	165 +/- 16		20.3 +/- 0.4
AMD RYZEN AI MAX PRO 390	138 +/- 4	9640 +/- 103	19.2 +/- 0.3
AMD EPYC 4545P 16-Core	111 +/- 6	12093 +/- 927	

Component	Prompt Processing [token/s]	Time To First Token [ms]	Token Generation [token/s]
AMD Ryzen 7 7840HS	79	17516 +/- 122	
AMD Ryzen AI 9 365	61 +/- 1	22242 +/- 499	12.1 +/- 0.3

- Model: Qwen2.5-14B-Instruct-Q4_K_M.gguf - Acceleration: CPU

Component	Prompt Processing [token/s]	Time To First Token [ms]	Token Generation [token/s]
AMD Ryzen 9 9950X 16-Core	125	11190 +/- 15	7.93 +/- 0.01
AMD Ryzen 9 9950X3D 16-Core	121 +/- 1	11460 +/- 85	7.83 +/- 0.07
AMD EPYC 4585PX 16-Core	116 +/- 15		6.19 +/- 0.03
AMD RYZEN AI MAX+ PRO 395	84 +/- 3	16158 +/- 467	11.11 +/- 0.27
AMD EPYC 4465P 12-Core	83	16717 +/- 23	6.53 +/- 0.01
AMD RYZEN AI MAX PRO 390	71 +/- 3	18981 +/- 334	10.52 +/- 0.27
AMD Ryzen AI 9 365			6.60 +/- 0.08
AMD Ryzen AI 9 HX 370			6.11 +/- 0.57

- Model: Qwen3-32B-Q4_K_M.gguf - Acceleration: CPU

Component	Prompt Processing [token/s]	Time To First Token [ms]	Token Generation [token/s]
AMD Ryzen 9 9950X 16-Core	53	26,414 +/- 49	3.60
AMD Ryzen 9 9950X3D 16-Core	52	26,991 +/- 160	3.57 +/- 0.02
AMD RYZEN AI MAX+ PRO 395	36 +/- 1	38,082 +/- 959	5.12 +/- 0.08
AMD RYZEN AI MAX PRO 390	30 +/- 1	43,996 +/- 273	4.85 +/- 0.07
AMD EPYC 4545P 16-Core	25 +/- 1	56,061 +/- 3,460	2.72 +/- 0.01
AMD Ryzen AI 9 HX 370			3.00 +/- 0.01

Performance

- https://scicomp.stackexchange.com/questions/36306/how-to-properly-calculate-cpu-and-gpu-flops-performance
- https://forums.developer.nvidia.com/t/how-to-measure-tensor-flops/292765
- https://docs.nvidia.com/deeplearning/performance/dl-performance-gpu-background/index.html
- https://medium.com/@shashanka_b_r/gpu-compute-performance-estimation-the-mathematical-foundation-behind-ai-hardware-benchmarks-7da221bcc9a4
- https://www.reddit.com/r/NintendoSwitch/comments/5urua3/explanation_of_flops_and_fp32_and_fp16/
- https://stackoverflow.com/questions/75853139/how-to-estimate-gpu-performance-using-clgetdeviceinfo
- https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/
- https://qwen.readthedocs.io/en/latest/getting_started/quantization_benchmark.html
- https://www.reddit.com/r/LocalLLaMA/comments/162pgx9/what_do_yall_consider_acceptable_tokens_per/
- https://www.reddit.com/r/ROCm/comments/1jfltc7/70b_llm_ts_speed_on_windows_rocm_using_24gb_rx/

Models

- https://arxiv.org/html/2409.12186v2#S2
- https://github.com/deepseek-ai/DeepSeek-V3

Benchmark

CPU benchmark

- https://openbenchmarking.org/result/2311131-NE-SIENAEPYC16%26sor%26sgm%3D1%26ppd_RVBZQyA4NTM0UA%3D5850%26ppd_RVBZQyA4NTM0UCAtIDE1NVc%3D5850%26ppd_RVBZQyA4NTM0UCAtIDIyNVcgUG93ZXI%3D5850%26ppd_RVBZQyA4NTM0UCAtIFBvd2Vy%3D5850%26ppd_RVBZQyA4NTM0UE4%3D6350%26ppd_RVBZQyA4NTM0UE4gLSBQb3dlcg%3D6350%26ppd_WGVvbiBQbGF0aW51bSA4NDY4%3D8024%26ppd_WGVvbiBQbGF0aW51bSA4NDY4IC0gNmMgUkFN%3D8324%26ppt%3DD%26sgm%3D1%26ppd_RVBZQyA4NTM0UA%3D5850%26ppd_RVBZQyA4NTM0UCAtIDE1NVc%3D5850%26ppd_RVBZQyA4NTM0UCAtIDIyNVcgUG93ZXI%3D5850%26ppd_RVBZQyA4NTM0UCAtIFBvd2Vy%3D5850%26ppd_RVBZQyA4NTM0UE4%3D6350%26ppd_RVBZQyA4NTM0UE4gLSBQb3dlcg%3D6350%26ppd_WGVvbiBQbGF0aW51bSA4NDY4%3D8324%26ppd_WGVvbiBQbGF0aW51bSA4NDY4IC0gNmMgUkFN%3D8024%26ppt%3DD
- https://openbenchmarking.org/s/2+x+AMD+EPYC+9175F+16-Core
- https://www.reddit.com/r/LocalLLaMA/comments/1iyztni/dual_9175f_amd_epyc_9005_a_new_trend/
- https://www.reddit.com/r/LocalLLaMA/comments/1jq13ik/comment/ml6hg70/?context=3
- https://www.reddit.com/r/threadripper/comments/1azmkvg/comparing_threadripper_7000_memory_bandwidth_for/

  > To get the best memory bandwidth, (theoretically) you should:
  >
  > - Increase FCLK for 8-channel configurations with 2 or 4 CCDs (7945WX, 7955WX, 7965WX, 7975WX),
  > - Use overclocked memory in all remaining Threadripper models,
  > - **For Epyc, purchase a motherboard with 12 memory slots and an Epyc 9004 processor with at least 8 CCDs. Fill all memory slots.**

- https://openbenchmarking.org/s/AMD+EPYC+8534PN+64-Core
- https://www.servethehome.com/amd-epyc-genoa-gaps-intel-xeon-in-stunning-fashion/3/
- https://superuser.com/questions/1815148/expected-results-of-a-stream-memory-bandwidth-benchmark

GPU benchmark

- https://llm-tracker.info/_TOORG/RTX-3090-vs-7900-XTX-Comparison

  > RTX 3090 vs 7900 XTX Comparison

- https://cprimozic.net/notes/posts/machine-learning-benchmarks-on-the-7900-xtx/

  > Machine Learning Benchmarks on the 7900 XTX

- https://espadrine.github.io/blog/posts/recomputing-gpu-performance.html
- https://www.reddit.com/r/LocalLLaMA/comments/191srof/amd_radeon_7900_xtxtx_inference_performance/
- https://embeddedllm.com/blog/vllm-now-supports-running-gguf-on-amd-radeon-gpu
- https://www.reddit.com/r/LocalLLaMA/comments/1atvxu2/current_state_of_training_on_amd_radeon_7900_xtx/
- https://www.tomshardware.com/reviews/gpu-hierarchy,4388.html#section-content-creation-gpu-benchmarks-rankings-2025
- https://public.tableau.com/app/profile/data.visualization6666/viz/MLCommons-Training_16993769118290/MLCommons-Training

Model benchmark

- https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/?official=true
- https://www.reddit.com/r/LocalLLaMA/comments/144rg6a/all_model_leaderboards_that_i_know/
- https://qwen.readthedocs.io/en/latest/getting_started/speed_benchmark.html

CPU + NPU benchmark

- https://www.reddit.com/r/LocalLLaMA/comments/1kmi3ra/amd_strix_halo_ryzen_ai_max_395_gpu_llm/
- https://old.chipsandcheese.com/2024/11/24/pushing-amds-infinity-fabric-to-its-limits/
- https://www.reddit.com/r/LocalLLaMA/comments/1ghvwsj/llamacpp_compute_and_memory_bandwidth_efficiency/
- https://www.reddit.com/r/LocalLLaMA/comments/1ghvwsj/comment/lv4sx1e/
- https://www.localscore.ai/
- https://wandb.ai/augmxnt/train-bench/reports/torchtune-vs-axolotl-vs-unsloth-Trainer-Comparison--Vmlldzo4MzU3NTAx
- https://github.com/underlines/awesome-ml/blob/master/llm-tools.md#benchmarking
- https://openbenchmarking.org/test/pts/llama-cpp-2.1.1

Performance

- https://semianalysis.com/2022/11/10/amd-genoa-detailed-architecture-makes/
- https://dev.to/maximsaplin/ddr5-speed-and-llm-inference-3cdn
- https://www.servethehome.com/guide-ddr-ddr2-ddr3-ddr4-and-ddr5-bandwidth-by-generation/
- https://www.hardware-corner.net/ddr-9000-ddr-8000-for-llm/
- https://www.techpowerup.com/review/ddr5-memory-performance-scaling-with-amd-zen-5/2.html
- https://blog.cloudflare.com/ddr4-memory-organization-and-how-it-affects-memory-bandwidth/#:~:text=Memory%20rank%20is%20a%20term,address%2C%20command%20and%20control%20signals.
- https://www.techpowerup.com/review/ddr5-memory-performance-scaling-with-amd-zen-5/22.html
- https://www.arukereso.hu/memoria-modul-c3577/kingston/fury-renegade-pro-32gb-ddr5-4800mhz-kf548r36rb-32-p943389321/

Motherboard

- https://www.asrockrack.com/general/productdetail.asp?Model=TURIN2D16-2T#Specifications
- https://pcpartpicker.com/products/motherboard/#s=41&mt=ddr5&sort=-rammax&E=2,7&m=7,8,46&f=2&c=167
- https://pangoly.com/en/browse/motherboard?extra=Socket%3AAM5;FormFactor%3AATX;Chipset%3AAMD%20X870E;M2Ports%3A2,8&sort=low
- https://www.arukereso.hu/alaplap-c3128/asrock/sienad8-2l2t-p1050062953/
- https://www.arukereso.hu/alaplap-c3128/asrock/sienad8-2l2t-p1050062953/
- https://www.gigabyte.com/Enterprise/Server-Motherboard/ME03-CE0-rev-10
- https://www.storagereview.com/news/asrock-rack-server-motherboards-with-amd-epyc-8004-series-processors-support-announced
- https://www.supermicro.com/en/products/motherboard/H13SSL-NT
- https://www.supermicro.com/en/products/motherboard/H13SSL-N
- https://www.asrockrack.com/general/productdetail.asp?Model=SIENAD8-2L2T#Specifications
- https://www.gigabyte.com/Enterprise/Server-Motherboard/ME03-CE1-rev-10
- https://www.reddit.com/r/homelab/comments/1aeio2n/amd_epyc_siena_cpus_6_memory_channels_and_asrock/
- https://www.senetic.hu/product/SIENAD8-2L2T
- https://www.asrockrack.com/general/productdetail.pl.asp?Model=GENOA2D24G-2L%2B#Specifications
- https://www.reddit.com/r/LocalLLaMA/comments/1e5g65f/i_found_a_nice_motherboard_for_an_imaginary_gpu/

SSD

- https://ezdiy-fab.com/products/m-2-nvme-ssd-pcie-4-0-adapter
- https://ezdiy-fab.com/products/m-2-nvme-ssd-pcie-4-0-adapter

PSU

GPU

NVIDIA GPU

- https://smicro.hu/nvidia-blackwell-5
- https://www.techpowerup.com/gpu-specs/geforce-rtx-5060-ti-16-gb.c4292
- https://bizon-tech.com/gpu-benchmarks/NVIDIA-RTX-4090-vs-NVIDIA-RTX-4500-Ada/637vs697
- https://www.arukereso.hu/videokartya-c3142/?st=RTX+5060+Ti
- https://www.tomshardware.com/pc-components/gpus/nvidia-geforce-rtx-5060-ti-16gb-review
- https://www.techpowerup.com/review/pny-geforce-rtx-5070-ti-epic-x-rgb-plus-oc/39.html
- https://www.arukereso.hu/videokartya-c3142/?st=RTX+5070+Ti
- https://www.tomshardware.com/pc-components/gpus/nvidia-rtx-pro-6000-blackwell-gpu-is-listed-for-usd8-565-at-us-retailer-26-percent-more-expensive-than-the-last-gen-rtx-6000-ada
- https://www.tomshardware.com/pc-components/gpus/rtx-5080-super-rumored-with-24gb-of-memory-same-10-752-cuda-cores-as-the-vanilla-variant-with-a-400w-tgp
- https://www.reddit.com/r/LocalLLaMA/comments/1lhd1j0/some_observations_using_the_rtx_6000_pro_blackwell/
- https://www.tomshardware.com/pc-components/gpus/nvidia-rtx-pro-6000-up-close-blackwell-rtx-workstation-max-q-workstation-and-server-variants-shown
- https://smicro.hu/nvidia-rtx-pro-6000-blackwell-max-q-ws-900-5g153-2200-000-4
- https://smicro.hu/pny-nvidia-rtx-pro-6000-server-edition-tcsrtxpro6000se-pb-4
- https://smicro.hu/nvidia-rtx-pro-6000-blackwell-server-edition-96gb-gddr7-ecc-passive-fhfl-600w-900-2g153-0000-000-4
- https://smicro.hu/pny-nvidia-rtx-pro-6000-blackwell-workstation-edition-vcnrtxpro6000-sb-4
- https://www.techpowerup.com/gpu-specs/rtx-pro-6000-blackwell-max-q.c4273
- https://www.techpowerup.com/gpu-specs/rtx-pro-6000-blackwell.c4272
- https://www.techpowerup.com/gpu-specs/rtx-pro-6000-blackwell-server.c4274

AMD GPU

- https://www.techpowerup.com/gpu-specs/asrock-rx-7900-xtx-creator.b11871
- https://www.techpowerup.com/gpu-specs/sapphire-ultimate-r7-250.b2751
- https://instinct.docs.amd.com/projects/amdgpu-docs/en/latest/gpu-partitioning/mi300a/overview.html
- https://www.reddit.com/r/LocalLLaMA/comments/1g0nrr0/experiences_on_running_7900xtx_to_run_llm_workload/
- https://www.newegg.com/xfx-speedster-merc310-rx-79xmercb9-radeon-rx-7900-xtx-24gb-graphics-card-triple-fans/p/N82E16814150878
- https://www.techpowerup.com/gpu-specs/

GPU link

- AMD GPUs support PCIe peer-to-peer out of the box, making it ideal for LLM training workloads that require tight coupling across multiple GPUs.
- https://c-payne.com/products/slimsas-sff-8654-to-sff-8654lp-low-profile-8i-cable-pcie-gen4

- https://en.wikipedia.org/wiki/List_of_interface_bit_rates

Case

Mini PC

- https://store.minisforum.com/products/elitemini-ai370
- https://minisforumpc.eu/en/products/ai-x1-pro-mini-pc?_pos=1&_psq=ai370&_ss=e&_v=1.0&variant=51875206496622
- https://minisforumpc.eu/en/products/ai-x1-pro-mini-pc?_pos=1&_psq=ai370&_ss=e&_v=1.0&variant=51875206496622
- https://minisforumpc.eu/en
- https://minisforumpc.eu/en/products/ai-x1-pro-mini-pc?variant=51875206496622
- https://www.hp.com/us-en/workstations/z2-mini-a.html
- https://www.techpowerup.com/333983/sapphire-develops-edge-ai-mini-pc-series-with-amd-ryzen-ai-300-targeting-gamers-and-creatives
- https://www.techpowerup.com/333983/sapphire-develops-edge-ai-mini-pc-series-with-amd-ryzen-ai-300-targeting-gamers-and-creatives
- https://www.reddit.com/r/LocalLLaMA/comments/1judxsq/gmktec_evox2_powered_by_ryzen_ai_max_395_to/
- https://frame.work/hu/en/products/framework-desktop-mainboard-amd-ryzen-ai-max-300-series?v=FRAFMK0006
- https://de.gmktec.com/en/products/gmktec-evo-x2-amd-ryzen%E2%84%A2-ai-max-395-mini-pc-1?variant=51106344992952

Mini PC + eGPU

- https://www.reddit.com/r/LocalLLaMA/comments/1kedbv7/ryzen_ai_max_395_a_gpu/

tinybox

- https://tinygrad.org/#tinybox
- https://docs.tinygrad.org/tinybox/
- https://tinycorp.myshopify.com/products/tinybox-red
- https://www.reddit.com/r/LocalLLaMA/comments/1gcn3w9/a_glance_inside_the_tinybox_pro_8_x_rtx_4090/
- https://news.ycombinator.com/item?id=41365637
- https://archive.md/RTvI8
- https://www.tomshardware.com/tech-industry/artificial-intelligence/ai-accelerator-tinybox-pro-goes-up-for-preorder-for-usd40-000-the-device-features-eight-rtx-4090s-and-two-amd-genoa-epyc-processors
- https://www.tweaktown.com/news/97110/tinycorps-new-tinybox-ai-system-amd-gpu-starts-at-15k-nvidia-25k/index.html
- https://www.tomshardware.com/pc-components/gpus/tinybox-ai-accelerator-now-available-starting-at-dollar15k-available-in-7900xtx-and-rtx-4090-variants
- George Hotz | AMD PC Build | tinygrad: building the new tinygrad computer | EPYC 7662 RX 7900 | ROCm: (Details) (Summary)
  - Date of stream 30 Apr 2023.
  - Excerpt from the technical summary:
    - Motherboard: ASRock Rack Rome a2t/BCM, 7 full bandwidth PCIe ports, $649 retail
    - CPU: AMD EPYC 7662 64-core processor, Gen 2 EPYC, $6,000 CPU from 2019, purchased for $800
    - RAM: 64GB Samsung (made in Philippines), 8GB sticks, single rank, 3200 MHz, ~$38 per stick, total around $300
    - GPUs: 2x AMD Radeon 7900 XTX (RDNA3 architecture), 24GB per GPU, 13 teraflops measured during testing, using PCIe extenders (16x PCIe4),

      - one GPU not connecting properly (likely due to longer extender), extender lengths: one 30cm, one 15cm
      - AMD consumer GPUs support peer-to-peer communication (NVIDIA consumer GPUs don't)
    - Storage: Samsung SSD 990 Pro 2TB NVMe
    - Power Supply: EVGA 1600 watt (primary), Dell server power supplies (2000W at 220V, limited to 1000W on US power) (backup), Dell 2400W power supplies that can do 1400W (alternative)
    - Cooling: Noctua fans (12 volt fans run at 5 volts), 3x 120mm fans, 1x 140mm fan, 2x 80mm fans
    - Network: Dual 10 gigabit network on PCIe
    - Software: Ubuntu Server 22.04 "Jammy Jellyfish", tinygrad, LLAMA model testing, ROCm stack for AMD GPUs
    - Approximately $5,000 total for the build
  - George Hotz | Exploring | Tenstorrent Blackhole on Arch Linux | tinycorp.myshopify.com | Part 2

    > Um so this is the tiny reimplementation of HLBC. 19 seconds. Okay. So, it's it's this 7900 XTX is 4x slower than a 5090, which yeah, it's about it's about a quarter the price. Yo, you know what? I'm half tempted to get rid of the discount on the red boxes. Like this is actually really usable.

PIM

- https://semiconductor.samsung.com/technologies/memory/pim/
- https://semiconductor.samsung.com/news-events/tech-blog/hbm-pim-cutting-edge-memory-technology-to-accelerate-next-generation-ai/
- https://www.techpowerup.com/278586/samsung-develops-industrys-first-high-bandwidth-memory-with-ai-processing-power
- https://www.techpowerup.com/328510/samsung-hopes-pim-memory-technology-can-replace-hbm-in-next-gen-ai-applications
- https://www.upmem.com/technology/
- https://www.upmem.com/
- https://sdk.upmem.com/master/00_ToolchainAtAGlance.html
- https://resources.sw.siemens.com/en-US/white-paper-innovative-upmem-pim-dram-requires-innovative-power-integrity-analysis/
- https://github.com/CMU-SAFARI/prim-benchmarks
- https://people.inf.ethz.ch/omutlu/pub/PrIM-UPMEM-Tutorial-Analysis-Benchmarking-SAFARI-Live-Seminar-2021-07-12-talk.pdf
- https://arxiv.org/html/2308.00846v3

Survey

- https://chatgpt.com/c/6830acbc-6258-800b-b8de-0b5a51a6ee13

- https://www.reddit.com/r/LocalLLaMA/comments/19crc6v/what_matters_cpuwise_for_gpu_inference/

- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=The%20biggest%20factor%20limiting%20performance,better%20suited%20for%20CPU%20inference

  > The biggest factor limiting performance during CPU inference is RAM memory bandwidth, and maximizing bandwidth directly leads to performance. This means that faster memory clock speed is preferred over lower latency, and platforms that support more memory channels, such as AMD Threadripper PRO or EPYCs, are better suited for CPU inference.

- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=In%20contrast%2C%20CPU%20inference%20uses,model%20on%20a%20capable%20GPU

  > In contrast, CPU inference uses the system RAM instead of the VRAM, making it much easier to run larger models that may not otherwise be able to be loaded into a system's VRAM. The major downside of this approach is that performance is going to be much lower, somewhere in the range of 10x to 100x slower, compared to running the same model on a capable GPU.

- https://news.ycombinator.com/item?id=37067933#:~:text=Anything%20with%2064GB%20of%20memory,That%20should%20generate%20faster%20than

- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=GPU%20inference%2C%20RAM%E2%80%99s%20primary%20use,RAM%E2%80%99s%20job%20is%20essentially%20done

  > During GPU inference, RAM's primary use is to facilitate loading a model's weights from storage into VRAM. For this reason, we recommend having at least as much RAM as VRAM in a system, but preferably 1.5-2x more RAM than VRAM. Attempting to load a model without sufficient RAM can fail if the capacity of the RAM + system page file is exceeded by the model, even if that model can fit on the available VRAM. However, once the model is loaded, then the RAM's job is essentially done.

- https://news.ycombinator.com/item?id=37067933#:~:text=Anything%20with%2064GB%20of%20memory,faster%20than%20you%20can%20read

- https://news.ycombinator.com/item?id=41046980#:~:text=match%20at%20L149%20A%20405B,disc%20and%20be%20unusably%20slow

- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=The%20biggest%20factor%20limiting%20performance,Threadripper%20PRO%20or%20EPYCs%2C%20are

  > Regardless of the inference method, storage does not play a significant role. Like RAM in the case of GPU inference, once a model is loaded, then there's not much for the storage to do. A drive's read performance does impact how quickly a model can be read and loaded into memory, so utilizing a fast NVMe drive to hold models' weights will help minimize the time spent loading into RAM or VRAM. But unless someone is frequently loading and testing a variety of different models, then this isn't likely to be much of a concern.

- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=In%20addition%20to%20strictly%20CPU,the%20greater%20the%20performance%20impact

  > In addition to strictly CPU or GPU approaches, there are inference libraries that support a hybrid method of inference utilizing both CPU/RAM and GPU/VRAM resources, most notably llama.cpp. This can be a good option for those who want to run a model that cannot fit entirely within their VRAM. Intuitively, this results in performance that lands somewhere between pure CPU and pure GPU inference. In this mode, the more of a model that has to be offloaded into system RAM, the greater the performance impact.

- https://www.reddit.com/r/LocalLLaMA/comments/1cj4det/llama_3_70b_instruct_works_surprisingly_well_on/

  > Llama 3 70b instruct works surprisingly well on 24gb VRAM cards

- https://medium.com/@markpalatucci/how-to-build-a-silent-multi-gpu-water-cooled-deep-learning-rig-for-under-10k-aefcdd1f96a5#:~:text=Power%20Supply%20,fan%20noise%20ramps%20considerably%20above

  > How to Build a Silent, Multi-GPU Water-Cooled Deep-Learning Rig for under $10k

- https://news.ycombinator.com/item?id=41046980#:~:text=Unsure%20if%20anyone%20has%20specific,1%20405b%20for%20roughly%20%2410k

  > One of the first forks in the road that you will encounter when starting with an LLM is whether to perform inference using the CPU or GPU. If you have a workstation with a graphics card released in the past five years or so, then it’s practically guaranteed that performing inference with your GPU will provide much better performance than if you were to use the CPU. However, especially with single-GPU configurations, the amount of VRAM that the GPU features (typically anywhere from 8GB to 48GB) is going to be the limiting factor with regard to which models that can be run via the GPU.

- https://modal.com/blog/how-much-vram-need-inference#:~:text=Let%E2%80%99s%20consider%204,70B

- https://huggingface.co/blog/llama31#:~:text=70B%20%20140%20GB%20,405%20GB%20%20203%20GB

- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=NVIDIA%20vs%20AMD%20GPU%20Performance,Capability

- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=support%20base%2C%20NVIDIA%20GPUs%20also,in%20terms%20of%20raw%20performance

- https://www.hardware-corner.net/guides/qwq-llm-rtx-3090-benchmark/

  > RTX 3090 Benchmarked Qwen QwQ AI Model - The 5000+ Token Context Test
  >
  > - Our most demanding test pushed the RTX 3090 near its memory limits with an 8,192-token context window allocation, though the actual context used was approximately 5,000 tokens. This configuration:

Metric	Value
Prompt Processing Time	8.5 seconds
Reasoning / Thinking Time	23 seconds
Generation Speed	19 tokens/second

- https://www.hardware-corner.net/amd-targets-faster-local-llms/

- https://www.hardware-corner.net/how-fast-ai-max-395-llm-20250317/

- https://www.hardware-corner.net/category/llm-news/page/2/

- https://www.hardware-corner.net/ai-playground-oss-arc-gpu-inference-20250418/

- https://www.hardware-corner.net/nvidia-rtx-5060-ti-16gb-spec-leaked/

- https://www.hardware-corner.net/dual-rtx-5060-ti-price-for-llm-build-20250415/

- https://www.hardware-corner.net/the-rtx-5060-ti-price-20250416/

- https://www.hardware-corner.net/nvidia-rtx-5060-ti-16gb-spec-leaked/

- https://www.hardware-corner.net/dual-rtx-5060-ti-price-for-llm-build-20250415/

- https://www.hardware-corner.net/the-rtx-5060-ti-price-20250416/

- https://www.hardware-corner.net/ai-playground-oss-arc-gpu-inference-20250418/

- https://www.nvidia.com/en-us/geforce/news/ultimate-guide-to-5060/

- https://www.hardware-corner.net/meta-releases-llama-4-what-hardware/

- https://www.hardware-corner.net/guides/install-llama-2-windows-pc/
- https://www.kitguru.net/channel/generaltech/joao-silva/amd-ryzen-ai-300-series-shows-impressive-llm-performance/
- https://i0.wp.com/timdettmers.com/wp-content/uploads/2023/01/gpu_recommendations.png?w=845&ssl=1
- https://timdettmers.com/
- https://timdettmers.com/2018/10/17/tpus-vs-gpus-for-transformers-bert/
- https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/
  - The Most Important GPU Specs for Deep Learning Processing Speed

    > Tensor Cores are most important, followed by memory bandwidth of a GPU, the cache hierarchy, and only then FLOPS of a GPU.

  - Tensor Cores

    > With Tensor Cores, we can perform a 4×4 matrix multiplication in one cycle. [...] Thus we reduce the matrix multiplication cost significantly from 504 cycles to 235 cycles via Tensor Cores. In this simplified case, the Tensor Cores reduced the cost of both shared memory access and FFMA operations.

  - Memory Bandwidth

    > Tensor Cores are very fast. So fast, in fact, that they are idle most of the time as they are waiting for memory to arrive from global memory. For example, during GPT-3-sized training, which uses huge matrices — the larger, the better for Tensor Cores — we have a Tensor Core TFLOPS utilization of about 45-65%, meaning that even for the large neural networks **about 50% of the time, Tensor Cores are idle**.

  - L2 Cache / Shared Memory / L1 Cache / Registers

    > A matrix memory tile in L2 cache is 3-5x faster than global GPU memory (GPU RAM), shared memory is ~7-10x faster than the global GPU memory, whereas the Tensor Cores' registers are ~200x faster than the global GPU memory.

- https://timdettmers.com/2018/12/16/deep-learning-hardware-guide/
- https://www.reddit.com/r/LocalLLaMA/comments/1erh260/2x_rtx_3090_threadripper_3970x_256gb_ram_llm/
- https://github.com/ggml-org/llama.cpp/discussions/11765
- https://www.pugetsystems.com/labs/hpc/exploring-hybrid-cpu-gpu-llm-inference/
- https://ipon.hu/shop/termek/amd-epyc-9115-26ghz-sp5-oem-100-000001552/2330933?aku=27dcaa5a946a5d25ecbc2b5ca46149b2
- https://hostbor.com/rtx-5060ti-vs-4060ti-comparison/
- https://blogs.nvidia.com/blog/ai-decoded-lm-studio/
- https://dev.to/maximsaplin/running-local-llms-cpu-vs-gpu-a-quick-speed-test-2cjn
- https://dev.to/maximsaplin/llamacpp-cpu-vs-gpu-shared-vram-and-inference-speed-3jpl
- https://dev.to/maximsaplin/fine-tuning-llm-on-a-laptop-vram-shared-memory-gpu-load-performance-4agj

- https://x-dev.pages.jsc.fz-juelich.de/2022/08/01/mi250-first-performances.html
- https://github.com/stas00/ml-engineering/blob/master/compute/accelerator/README.md
- https://docs.nvidia.com/nim/large-language-models/latest/supported-models.html
- https://www.phoronix.com/review/intel-sapphirerapids-avx512/2
- https://community.amd.com/t5/ai/integration-ascendant-exploring-the-amd-ryzen-ai-max-pro-series/ba-p/753606
- https://www.reddit.com/r/LocalLLaMA/comments/15rwe7t/the_llm_gpu_buying_guide_august_2023/
- https://www.pugetsystems.com/labs/articles/llm-inference-consumer-gpu-performance/
- https://www.hyperstack.cloud/technical-resources/tutorials/how-to-choose-the-right-gpu-for-llm-a-practical-guide
- https://www.reddit.com/r/LocalLLaMA/comments/1iyztni/dual_9175f_amd_epyc_9005_a_new_trend/
- https://www.reddit.com/r/LocalLLaMA/comments/1k57b1o/what_workstationrack_should_i_buy_for_offline_llm/
- https://chipsandcheese.com/p/amds-cdna-3-compute-architecture

Software

Software Optimization

Efficient libraries (e.g., BLAS, oneDNN) and threading can significantly impact throughput.

AMD CPU

- https://www.amd.com/en/developer/resources/technical-articles/zendnn-5-0-supercharge-ai-on-amd-epyc-server-cpus.html
- https://github.com/ggml-org/llama.cpp/issues/11744
- https://github.com/ggml-org/llama.cpp/discussions/11733

AMD GPU

- https://github.com/ROCm/flash-attention/tree/howiejay/navi_support
- https://github.com/ROCm/flash-attention
- https://llm-tracker.info/howto/AMD-GPUs
- https://www.youtube.com/watch?v=yCCoQ72DBpM

NVIDIA GPU

- https://developer.nvidia.com/nccl

Multi-GPU

- https://medium.com/@geronimo7/llms-multi-gpu-inference-with-accelerate-5a8333e4c5db
- https://docs.vllm.ai/en/v0.8.0/serving/distributed_serving.html

Library

- https://nn-512.com/
- https://github.com/explosion/thinc/tree/main

Framework

- llamafile - Lets you distribute and run LLMs with a single file.
  - LLaMA 3.2 3B Instruct - llamafile -
    - A large language model small enough to run on most computers with 8GB+ of RAM.

tinygrad

- https://tinygrad.org/#tinygrad
- https://news.ycombinator.com/item?id=33462337
- https://www.phoronix.com/forums/forum/linux-graphics-x-org-drivers/open-source-amd-linux/1519271-tiny-corp-nearing-completely-sovereign-compute-stack-for-amd-gpus-with-tinygrad/page2
- https://www.youtube.com/@geohotarchive/videos
- https://docs.tinygrad.org/showcase/
- https://github.com/tinygrad/tinygrad
- https://www.youtube.com/watch?v=Xtws3-Pk69o&list=PLzFUMGbVxlQsh0fFZ2QKOBY25lz04A3hi
- Phoronix: Tiny Corp Closing In On "Completely Sovereign" Compute Stack For AMD GPUs With Tinygrad

> By the end of this year, I'm confident we'll have end to end perf similar to a 4090 in PyTorch on the 7900XTX. But this is a long journey where each piece needs to work together. End of the year, you can hold me to that.

TextSynth Server

- https://bellard.org/ts_server/
- https://bellard.org/ts_server/ts_server.html
- https://textsynth.com/documentation.html#embeddings
- https://textsynth.com/documentation.html#engines
- https://huggingface.co/fbellard/ts_server/tree/main
- https://github.com/sgl-project/sglang
- https://docs.sglang.ai/index.html
- https://dlib.net/ml.html#add_layer
- https://github.com/sony/nnabla
- https://www.nomic.ai/gpt4all
- https://www.phoronix.com/news/Red-Hat-llm-d-AI-LLM-Project
- https://github.com/unslothai/unsloth
- https://moondream.ai/
- https://llava-vl.github.io/
- https://github.com/bitsandbytes-foundation/bitsandbytes

Text

Visual analysis and generation

- https://lightning.ai/pages/community/serve-stable-diffusion-three-times-faster/
- Moondream - Open-source visual language model that understands images using simple text prompts. Fast and wildly capable.
- LLaVA - Large Language and Vision Assistant

Voice

- https://huggingface.co/mistralai/Voxtral-Mini-3B-2507
- https://huggingface.co/mistralai/Voxtral-Mini-3B-2507#vllm-recommended
- https://mistral.ai/news/voxtral
- https://github.com/mozilla/DeepSpeech
- Piper - A fast, local neural TTS optimized for the Raspberry Pi 4.

OCR

- https://medium.com/@fuelyourdigital/how-to-extract-data-from-a-graph-image-with-chatgpt-4-6a26351e4227
- https://apps.automeris.io/wpd4/
- https://automeris.io/
- https://automeris.io/docs/
- https://zertrin.org/webplotdigitizer/
- https://github.com/automeris-io/WebPlotDigitizer
- https://web.eecs.utk.edu/~dcostine/personal/PowerDeviceLib/DigiTest/index.html
- https://www.colliseum.net/WebPlot/
- https://alternativeto.net/software/graphclick/?platform=windows
- https://alternativeto.net/software/graphclick/
- https://academia.stackexchange.com/questions/7671/software-for-extracting-data-from-a-graph-without-having-to-click-on-every-singl
- https://www.im2graph.co.il/
- https://github.com/Cvrane/ChartReader
- http://www.graphreader.com/
- https://plotdigitizer.com/
- https://www.microsoft.com/en-us/research/publication/chartocr-data-extraction-from-charts-images-via-a-deep-hybrid-framework/
- https://mathematica.stackexchange.com/questions/102362/reconstruct-a-graph-from-an-image-curved-edges-and-edge-crossings
- https://mathematica.stackexchange.com/questions/102262/how-to-generate-a-graph-from-a-picture-of-a-graph?lq=1
- https://www.geeksforgeeks.org/image-reconstruction-using-singular-value-decomposition-svd-in-python/
- https://medium.com/@pranjallk1995/pca-for-image-reconstruction-from-scratch-cf4a787c1e36
- http://www.fmwconcepts.com/imagemagick/textcleaner/index.php
- https://willus.com/blog.shtml?tesseract_accuracy
- https://www.zdnet.com/article/the-next-decade-in-ai-gary-marcus-four-steps-towards-robust-artificial-intelligence/
- https://github.com/zacharywhitley/awesome-ocr
- https://en.wikipedia.org/wiki/OCRopus
- https://en.wikipedia.org/wiki/Tesseract_(software)
- https://upmath.me/

Embedding

- https://radimrehurek.com/gensim/auto_examples/core/run_core_concepts.html
- https://github.com/piskvorky/gensim
- https://radimrehurek.com/gensim/#
- https://en.wikipedia.org/wiki/ELMo
- https://en.wikipedia.org/wiki/Cosine_similarity
- https://en.wikipedia.org/wiki/Word2vec
- https://en.wikipedia.org/wiki/FastText

- https://en.wikipedia.org/wiki/Gensim

Papers

- https://arxiv.org/html/2505.07203v1
- https://github.com/d893a/localscore-leaderboard
- https://github.com/d893a/local-ml-models/blob/main/local_ml_models.md
- George Hotz blog - Nobody Profits

  > The best outcome of AI is if it delivers huge amounts of value to society but no profit to anyone.

  > Spin up open source projects in every sector to eliminate all the capturable value. This is what I'm trying to do with comma.ai and tinygrad.

Links:

Local LLM

Performance

- https://scicomp.stackexchange.com/questions/36306/how-to-properly-calculate-cpu-and-gpu-flops-performance
- https://www.reddit.com/r/NintendoSwitch/comments/5urua3/explanation_of_flops_and_fp32_and_fp16/
- https://www.reddit.com/r/LocalLLaMA/comments/162pgx9/what_do_yall_consider_acceptable_tokens_per/
- https://huggingface.co/docs/transformers/model_memory_anatomy
- https://en.wikipedia.org/wiki/Floating_point_operations_per_second

Memory Performance

- https://dev.to/maximsaplin/ddr5-speed-and-llm-inference-3cdn
- https://www.servethehome.com/guide-ddr-ddr2-ddr3-ddr4-and-ddr5-bandwidth-by-generation/
- https://www.hardware-corner.net/ddr-9000-ddr-8000-for-llm/
- https://www.techpowerup.com/review/ddr5-memory-performance-scaling-with-amd-zen-5/2.html
- https://www.techpowerup.com/review/ddr5-memory-performance-scaling-with-amd-zen-5/22.html
- https://www.techpowerup.com/forums/threads/infinity-fabric-bandwidth-vs-ram-bandwidth.324862/

CPU Performance

- https://github.com/ggml-org/llama.cpp/discussions/11733
- https://justine.lol/matmul/
- https://www.reddit.com/r/LocalLLaMA/comments/1bt8kc9/comparing_the_performance_of_epyc_9374f_and/
- https://blog.cloudflare.com/gen-12-servers/
- https://old.chipsandcheese.com/2024/11/24/pushing-amds-infinity-fabric-to-its-limits/
- https://www.reddit.com/r/LocalLLaMA/comments/19crc6v/what_matters_cpuwise_for_gpu_inference/

GPU performance

- https://forums.developer.nvidia.com/t/how-to-measure-tensor-flops/292765
- https://docs.nvidia.com/deeplearning/performance/dl-performance-gpu-background/index.html
- https://medium.com/@shashanka_b_r/gpu-compute-performance-estimation-the-mathematical-foundation-behind-ai-hardware-benchmarks-7da221bcc9a4
- https://stackoverflow.com/questions/75853139/how-to-estimate-gpu-performance-using-clgetdeviceinfo

- https://developer.nvidia.com/blog/mastering-llm-techniques-inference-optimization/
- https://www.youtube.com/watch?v=LSQL7c29arM
- https://www.reddit.com/r/LocalLLaMA/comments/1g0nrr0/experiences_on_running_7900xtx_to_run_llm_workload/
- https://www.reddit.com/r/ROCm/comments/1jfltc7/70b_llm_ts_speed_on_windows_rocm_using_24gb_rx/

Benchmark

CPU benchmark

- https://openbenchmarking.org/result/2311131-NE-SIENAEPYC16%26sor%26sgm%3D1%26ppd_RVBZQyA4NTM0UA%3D5850%26ppd_RVBZQyA4NTM0UCAtIDE1NVc%3D5850%26ppd_RVBZQyA4NTM0UCAtIDIyNVcgUG93ZXI%3D5850%26ppd_RVBZQyA4NTM0UCAtIFBvd2Vy%3D5850%26ppd_RVBZQyA4NTM0UE4%3D6350%26ppd_RVBZQyA4NTM0UE4gLSBQb3dlcg%3D6350%26ppd_WGVvbiBQbGF0aW51bSA4NDY4%3D8024%26ppd_WGVvbiBQbGF0aW51bSA4NDY4IC0gNmMgUkFN%3D8324%26ppt%3DD%26sgm%3D1%26ppd_RVBZQyA4NTM0UA%3D5850%26ppd_RVBZQyA4NTM0UCAtIDE1NVc%3D5850%26ppd_RVBZQyA4NTM0UCAtIDIyNVcgUG93ZXI%3D5850%26ppd_RVBZQyA4NTM0UCAtIFBvd2Vy%3D5850%26ppd_RVBZQyA4NTM0UE4%3D6350%26ppd_RVBZQyA4NTM0UE4gLSBQb3dlcg%3D6350%26ppd_WGVvbiBQbGF0aW51bSA4NDY4%3D8324%26ppd_WGVvbiBQbGF0aW51bSA4NDY4IC0gNmMgUkFN%3D8024%26ppt%3DD
- https://openbenchmarking.org/s/2+x+AMD+EPYC+9175F+16-Core
- https://www.reddit.com/r/LocalLLaMA/comments/1iyztni/dual_9175f_amd_epyc_9005_a_new_trend/
- https://www.reddit.com/r/LocalLLaMA/comments/1jq13ik/comment/ml6hg70/?context=3
- https://www.reddit.com/r/threadripper/comments/1azmkvg/comparing_threadripper_7000_memory_bandwidth_for/
- https://openbenchmarking.org/s/AMD+EPYC+8534PN+64-Core
- https://www.servethehome.com/amd-epyc-genoa-gaps-intel-xeon-in-stunning-fashion/3/
- https://superuser.com/questions/1815148/expected-results-of-a-stream-memory-bandwidth-benchmark

GPU benchmark

- https://cprimozic.net/notes/posts/machine-learning-benchmarks-on-the-7900-xtx/
- https://espadrine.github.io/blog/posts/recomputing-gpu-performance.html
- https://www.reddit.com/r/LocalLLaMA/comments/191srof/amd_radeon_7900_xtxtx_inference_performance/
- https://embeddedllm.com/blog/vllm-now-supports-running-gguf-on-amd-radeon-gpu
- https://www.reddit.com/r/LocalLLaMA/comments/1atvxu2/current_state_of_training_on_amd_radeon_7900_xtx/
- https://www.tomshardware.com/reviews/gpu-hierarchy,4388.html#section-content-creation-gpu-benchmarks-rankings-2025
- https://public.tableau.com/app/profile/data.visualization6666/viz/MLCommons-Training_16993769118290/MLCommons-Training
- https://llm-tracker.info/RTX-PRO-6000
- https://github.com/AUGMXNT/speed-benchmarking/tree/main/nvfp4
- https://www.hardware-corner.net/guides/qwq-llm-rtx-3090-benchmark/
- https://llm-tracker.info/_TOORG/RTX-3090-vs-7900-XTX-Comparison

NVIDIA DGX Spark

- https://github.com/ggml-org/llama.cpp/discussions/16578

Model benchmark

- https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/?official=true
- https://www.reddit.com/r/LocalLLaMA/comments/144rg6a/all_model_leaderboards_that_i_know/

- https://qwen.readthedocs.io/en/latest/getting_started/speed_benchmark.html

CPU + NPU benchmark

- https://www.reddit.com/r/LocalLLaMA/comments/1kmi3ra/amd_strix_halo_ryzen_ai_max_395_gpu_llm/
- https://llm-tracker.info/_TOORG/Strix-Halo
- https://www.reddit.com/r/LocalLLaMA/comments/1ghvwsj/llamacpp_compute_and_memory_bandwidth_efficiency/
- https://www.reddit.com/r/LocalLLaMA/comments/1ghvwsj/comment/lv4sx1e/
- https://www.localscore.ai/
- https://wandb.ai/augmxnt/train-bench/reports/torchtune-vs-axolotl-vs-unsloth-Trainer-Comparison--Vmlldzo4MzU3NTAx
- https://github.com/underlines/awesome-ml/blob/master/llm-tools.md#benchmarking
- https://openbenchmarking.org/test/pts/llama-cpp-2.1.1
- https://idp-leaderboard.org/
- https://www.reddit.com/r/DeepSeek/comments/1lhzskj/2x_nvidia_rtx_6000_blackwell_gpus_in_my_ai/
- https://www.reddit.com/r/LocalLLaMA/comments/1o6u5o4/comment/njl2no6/?context=3
- https://kyuz0.github.io/amd-strix-halo-toolboxes/

Models

- https://arxiv.org/html/2409.12186v2#S2
- https://github.com/deepseek-ai/DeepSeek-V3
- https://reka.ai/
- https://qwen.readthedocs.io/en/latest/getting_started/quantization_benchmark.html
- https://github.com/ikawrakow/ik_llamafile
- https://huggingface.co/blog/llama31#:~:text=70B%20%20140%20GB%20,405%20GB%20%20203%20GB

Survey

- https://www.hardware-corner.net/amd-targets-faster-local-llms/
- https://www.hardware-corner.net/how-fast-ai-max-395-llm-20250317/
- https://www.hardware-corner.net/category/llm-news/page/2/
- https://www.hardware-corner.net/ai-playground-oss-arc-gpu-inference-20250418/
- https://www.hardware-corner.net/nvidia-rtx-5060-ti-16gb-spec-leaked/
- https://www.hardware-corner.net/dual-rtx-5060-ti-price-for-llm-build-20250415/
- https://www.hardware-corner.net/the-rtx-5060-ti-price-20250416/
- https://www.hardware-corner.net/nvidia-rtx-5060-ti-16gb-spec-leaked/
- https://www.hardware-corner.net/dual-rtx-5060-ti-price-for-llm-build-20250415/
- https://www.hardware-corner.net/the-rtx-5060-ti-price-20250416/
- https://www.hardware-corner.net/ai-playground-oss-arc-gpu-inference-20250418/
- https://www.nvidia.com/en-us/geforce/news/ultimate-guide-to-5060/
- https://www.hardware-corner.net/meta-releases-llama-4-what-hardware/
- https://www.hardware-corner.net/guides/install-llama-2-windows-pc/
- https://www.kitguru.net/channel/generaltech/joao-silva/amd-ryzen-ai-300-series-shows-impressive-llm-performance/
- https://i0.wp.com/timdettmers.com/wp-content/uploads/2023/01/gpu_recommendations.png?w=845&ssl=1
- https://timdettmers.com/
- https://timdettmers.com/2018/10/17/tpus-vs-gpus-for-transformers-bert/
- https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/
- https://timdettmers.com/2018/12/16/deep-learning-hardware-guide/

- https://www.reddit.com/r/LocalLLaMA/comments/1erh260/2x_rtx_3090_threadripper_3970x_256gb_ram_llm/
- https://github.com/ggml-org/llama.cpp/discussions/11765
- https://www.pugetsystems.com/labs/hpc/exploring-hybrid-cpu-gpu-llm-inference/
- https://ipon.hu/shop/termek/amd-epyc-9115-26ghz-sp5-oem-100-000001552/2330933?aku=27dcaa5a946a5d25ecbc2b5ca46149b2
- https://hostbor.com/rtx-5060ti-vs-4060ti-comparison/
- https://blogs.nvidia.com/blog/ai-decoded-lm-studio/
- https://dev.to/maximsaplin/running-local-llms-cpu-vs-gpu-a-quick-speed-test-2cjn
- https://dev.to/maximsaplin/llamacpp-cpu-vs-gpu-shared-vram-and-inference-speed-3jpl
- https://dev.to/maximsaplin/fine-tuning-llm-on-a-laptop-vram-shared-memory-gpu-load-performance-4agj
- https://x-dev.pages.jsc.fz-juelich.de/2022/08/01/mi250-first-performances.html
- https://github.com/stas00/ml-engineering/blob/master/compute/accelerator/README.md
- https://docs.nvidia.com/nim/large-language-models/latest/supported-models.html
- https://www.phoronix.com/review/intel-sapphirerapids-avx512/2
- https://community.amd.com/t5/ai/integration-ascendant-exploring-the-amd-ryzen-ai-max-pro-series/ba-p/753606
- https://www.reddit.com/r/LocalLLaMA/comments/15rwe7t/the_llm_gpu_buying_guide_august_2023/
- https://www.pugetsystems.com/labs/articles/llm-inference-consumer-gpu-performance/
- https://www.hyperstack.cloud/technical-resources/tutorials/how-to-choose-the-right-gpu-for-llm-a-practical-guide
- https://www.reddit.com/r/LocalLLaMA/comments/1iyztni/dual_9175f_amd_epyc_9005_a_new_trend/
- https://www.reddit.com/r/LocalLLaMA/comments/1k57b1o/what_workstationrack_should_i_buy_for_offline_llm/
- https://chipsandcheese.com/p/amds-cdna-3-compute-architecture

Hardware

CPU

- https://www.amd.com/en/products/specifications/processors.html
- https://www.techpowerup.com/cpu-specs/
- https://www.techpowerup.com/327388/amd-granite-ridge-zen-5-processor-annotated

AMD EPYC

AMD EPYC 9004

- https://semianalysis.com/2022/11/10/amd-genoa-detailed-architecture-makes/
- https://www.cpubenchmark.net/high_end_cpus.html

Memory

- https://blog.cloudflare.com/ddr4-memory-organization-and-how-it-affects-memory-bandwidth/#:~:text=Memory%20rank%20is%20a%20term,address%2C%20command%20and%20control%20signals.
- https://www.servethehome.com/memory-bandwidth-per-core-and-per-socket-for-intel-xeon-and-amd-epyc/

CXL Memory extension

- https://en.wikipedia.org/wiki/Compute_Express_Link
- https://www.gigabyte.com/PC-Accessory/AI-TOP-CXL-R5X4/support#dl

AMD Ryzen Overclocked Memory

- https://www.amd.com/en/products/processors/ryzen-compatible-memory.html

SSD

- https://ezdiy-fab.com/products/m-2-nvme-ssd-pcie-4-0-adapter
- https://www.techpowerup.com/ssd-specs/

PSU

- https://en.wikipedia.org/wiki/80_Plus
- https://www.clearesult.com/80plus/certified-psus/all-certified-psus?page=13
- https://www.asus.com/motherboards-components/power-supply-units/workstation/asus-pro-ws-3000p/
- https://www.asus.com/motherboards-components/power-supply-units/workstation/asus-pro-ws-2200p/
- https://seasonic.com/atx3-prime-px-2200/

GPU

NVIDIA

- https://www.techpowerup.com/gpu-specs/geforce-rtx-5060-ti-16-gb.c4292
- https://www.tomshardware.com/pc-components/gpus/nvidia-geforce-rtx-5060-ti-16gb-review
- https://www.techpowerup.com/review/pny-geforce-rtx-5070-ti-epic-x-rgb-plus-oc/39.html
- https://www.tomshardware.com/pc-components/gpus/rtx-5080-super-rumored-with-24gb-of-memory-same-10-752-cuda-cores-as-the-vanilla-variant-with-a-400w-tgp
- https://www.reddit.com/r/LocalLLaMA/comments/1lhd1j0/some_observations_using_the_rtx_6000_pro_blackwell/
- https://www.tomshardware.com/pc-components/gpus/nvidia-rtx-pro-6000-up-close-blackwell-rtx-workstation-max-q-workstation-and-server-variants-shown
- https://www.tomshardware.com/pc-components/gpus/nvidia-rtx-pro-6000-blackwell-gpu-is-listed-for-usd8-565-at-us-retailer-26-percent-more-expensive-than-the-last-gen-rtx-6000-ada
- https://smicro.hu/nvidia-blackwell-5
- https://smicro.hu/nvidia-rtx-pro-6000-blackwell-max-q-ws-900-5g153-2200-000-4
- https://smicro.hu/pny-nvidia-rtx-pro-6000-server-edition-tcsrtxpro6000se-pb-4
- https://smicro.hu/nvidia-rtx-pro-6000-blackwell-server-edition-96gb-gddr7-ecc-passive-fhfl-600w-900-2g153-0000-000-4
- https://smicro.hu/pny-nvidia-rtx-pro-6000-blackwell-workstation-edition-vcnrtxpro6000-sb-4
- https://www.techpowerup.com/gpu-specs/rtx-pro-6000-blackwell-max-q.c4273
- https://www.techpowerup.com/gpu-specs/rtx-pro-6000-blackwell.c4272
- https://www.techpowerup.com/gpu-specs/rtx-pro-6000-blackwell-server.c4274
- https://www.techpowerup.com/339358/nvidia-to-debut-geforce-rtx-50-series-super-gpus-by-christmas

AMD

- https://www.techpowerup.com/gpu-specs/asrock-rx-7900-xtx-creator.b11871
- https://instinct.docs.amd.com/projects/amdgpu-docs/en/latest/gpu-partitioning/mi300a/overview.html
- https://www.newegg.com/xfx-speedster-merc310-rx-79xmercb9-radeon-rx-7900-xtx-24gb-graphics-card-triple-fans/p/N82E16814150878
- https://www.asrock.com/Graphics-Card/AMD/Radeon%20AI%20PRO%20R9700%20Creator%2032GB/index.asp
- https://www.techpowerup.com/gpu-specs/

GPU link

- https://c-payne.com/products/slimsas-sff-8654-to-sff-8654lp-low-profile-8i-cable-pcie-gen4
- https://en.wikipedia.org/wiki/List_of_interface_bit_rates

Chassis

- https://www.supermicro.com/en/products/chassis/4U/747/SC747BTQ-R2K04B
- https://www.youtube.com/watch?v=VrTHwN6OKG0
- https://www.corsair.com/us/en/explorer/diy-builder/cases/corsair-9000d-rgb-airflow/

Platform

Mini PC / Mobile

CPU

- https://www.techpowerup.com/cpu-specs/ryzen-ai-max-pro-395.c3998
- https://www.amd.com/en/products/processors/laptop/ryzen-pro/ai-max-pro-300-series/amd-ryzen-ai-max-plus-pro-395.html
- https://www.amd.com/en/blogs/2025/amd-ryzen-ai-max-395-processor-breakthrough-ai-.html
- https://www.tomshardware.com/pc-components/cpus/more-affordable-strix-halo-model-emerges-early-ryzen-ai-max-385-geekbench-result-reveals-an-eight-core-option
- https://store.minisforum.com/products/elitemini-ai370
- https://minisforumpc.eu/en/products/ai-x1-pro-mini-pc?_pos=1&_psq=ai370&_ss=e&_v=1.0&variant=51875206496622
- https://minisforumpc.eu/en
- https://minisforumpc.eu/en/products/ai-x1-pro-mini-pc?variant=51875206496622
- https://www.hp.com/us-en/workstations/z2-mini-a.html
- https://www.techpowerup.com/333983/sapphire-develops-edge-ai-mini-pc-series-with-amd-ryzen-ai-300-targeting-gamers-and-creatives
- https://www.techpowerup.com/333983/sapphire-develops-edge-ai-mini-pc-series-with-amd-ryzen-ai-300-targeting-gamers-and-creatives
- https://www.reddit.com/r/LocalLLaMA/comments/1judxsq/gmktec_evox2_powered_by_ryzen_ai_max_395_to/
- https://frame.work/hu/en/products/framework-desktop-mainboard-amd-ryzen-ai-max-300-series?v=FRAFMK0006
- https://de.gmktec.com/en/products/gmktec-evo-x2-amd-ryzen%E2%84%A2-ai-max-395-mini-pc-1?variant=51106344992952
- https://www.jeffgeerling.com/blog/2025/i-clustered-four-framework-mainboards-test-huge-llms
- https://github.com/geerlingguy/beowulf-ai-cluster
- https://deskpi.com/products/deskpi-rackmate-t1-black-rackmount-10-inch-8u-server-cabinet-for-network-servers-audio-and-video-equipment
- https://www.arukereso.hu/pc-konfiguracio-c3083/hp/z2-mini-g1a-b34l0es-p1218977497/
- https://www.mysoft.hu/details.aspx?pn=B34L0ES

Mini PC + eGPU

- https://www.reddit.com/r/LocalLLaMA/comments/1kedbv7/ryzen_ai_max_395_a_gpu/
- https://www.techpowerup.com/339051/peladn-unveils-link-s-3-its-first-thunderbolt-5-egpu-dock
- https://peladn.com/products/graphics-card-docking-station-1
- https://youtu.be/L-xgMQ-7lW0?t=1159
- https://www.razer.com/au-en/gaming-egpus/razer-core-x-v2?irclickid=0BfRkSytBxycRN2xAVT3Xwl8Ukp2hqTX4RbQ0c0&irgwc=1&utm_sharedid=&cid=Truong%20Nguyen-affiliate
- https://www.amazon.de/-/en/Thunderbolt-External-Station-1000Mbps-10-13-4-black/dp/B0DG83VNHQ/259-8824975-8309746?pd_rd_w=wTvag&content-id=amzn1.sym.5ce46963-b829-421e-9856-1a7676e820a4:amzn1.symc.1042562f-235b-4049-91eb-05d433a5d976&pf_rd_p=5ce46963-b829-421e-9856-

1a7676e820a4&pf_rd_r=PBFBFMCR6NZJA6FG4QKW&pd_rd_wg=dr337&pd_rd_r=d4283189-fc22-4d13-9840-6785a0c7826d&pd_rd_i=B0DG83VNHQ&psc=1

- https://aoostar.com/products/aoostar-ag01-egpu-dock-with-oculink-port-built-in-huntkey-400w-power-supply-supports-tgx-interface-hot-swap
- https://www.amazon.de/-/en/Graphics-Docking-External-Compatible-Charging/dp/B0CG2CXM98?crid=130F5OZ3VAMY5&dib=eyJ2IjoiMSJ9.Z-W5zXTF6bwPftg1TnwwCEppFLWq1SgC2VPFlQE_b7N2SscXk7QPyEh5eaYFHKGHNgBNh64n9oXXOHC8foGLeDqgBHAe8dSDFy0SjH3duw0mMPQcqyYmbJaNqN8062Hm9eYZsVbNNPcdNv3UNmwfmAP5vHGTiqaUlXG5H9pfyBE73cI6QfEc2VUc3ucZDLQlMidj_wBoiA0VJAGk4wVm_ClVAvvACy06sjtpjruykpLuU-fIUOOxnUPrFPzXtHqwy8_IG9vYp-z9frDIQ9wDFAn22qQkdkQcK2PCl955FYA.Z_lvjQW60H7Wn6yfmxF7EGQxooLoF5PbssYJYMDZmWw&dib_tag=se&keywords=Th3p4g3&qid=1755212846&s=computers&sprefix=th3p4g3%2Ccomputers%2C179&sr=1-1
- https://www.amazon.de/-/en/TH3P4G3-Enclosure-Thunderbolt-Compatible-SFX-FlEX/dp/B0F4MF7QFC?crid=130F5OZ3VAMY5&dib=eyJ2IjoiMSJ9.Z-W5zXTF6bwPftg1TnwwCEppFLWq1SgC2VPFlQE_b7N2SscXk7QPyEh5eaYFHKGHNgBNh64n9oXXOHC8foGLeDqgBHAe8dSDFy0SjH3duw0mMPQcqyYmbJaNqN8062Hm9eYZsVbNNPcdNv3UNmwfmAP5vHGTiqaUlXG5H9pfyBE73cI6QfEc2VUc3ucZDLQlMidj_wBoiA0VJAGk4wVm_ClVAvvACy06sjtpjruykpLuU-fIUOOxnUPrFPzXtHqwy8_IG9vYp-z9frDIQ9wDFAn22qQkdkQcK2PCl955FYA.Z_lvjQW60H7Wn6yfmxF7EGQxooLoF5PbssYJYMDZmWw&dib_tag=se&keywords=Th3p4g3&qid=1755212846&s=computers&sprefix=th3p4g3%2Ccomputers%2C179&sr=1-2
- https://www.amazon.de/-/en/Thunderbolt-External-Station-1000Mbps-10-13-4-black/dp/B0DG83VNHQ?crid=235D4PTBNX5ZP&dib=eyJ2IjoiMSJ9.g7sYoYi_MNEEYthOBdFzvumHqRqTfWjH4dfqSzGH2hbROrgPqZXhV6TBwAvDdUFzKk5nfuZ_bXRDDl8pNEoxXUoiDO3EQIGLotfrHmvq7qKGy0GNkZqjqlgqf0BVQskNg7UQyRTBrc7zJ0fEoDEMy1SlJRPxMahQWX_ihmGgVZJvKXTLWUlLWghfGY31a-15PvXbY8qAYPhvdihK_PX6VEgCLL7mVhLIL1sMqyeuk4R52KSv8bwsTO7WeFJvSrE6OHCJguYnlOM7ub3D8R4PKWCKIYfE0s3nOY4xlHP6NB4.v8M3VLe4bRHL8htnNb9envDNm65sUpmlOdu_nR8NE1Q&dib_tag=se&keywords=peladn+gpu+dock&qid=1755213168&s=computers&sprefix=peladn+gpu+dock%2Ccomputers%2C97&sr=1-17
- https://www.youtube.com/watch?v=wsmFcbWMCDU
- https://www.youtube.com/watch?v=FFedqJFIQwg
- https://www.youtube.com/watch?v=rQ8XR9xhVBU
- https://egpu.io/forums/builds/2021-14-hp-elitebook-840-g8-11th4cg-rtx-4070-32gbps-tb4-adt-link-r43sg-tb3-win11/
- https://www.razer.com/au-en/gaming-egpus/razer-core-x-v2?irclickid=0BfRkSytBxycRN2xAVT3Xwl8Ukp2m3zO4RbQ0c0&irgwc=1&utm_sharedid=&cid=Truong%20Nguyen-affiliate
- https://hardverapro.hu/apro/lenovo_legion_booststation_thunderbolt_egpu_hdd_ne_3/friss.html
- https://www.techradar.com/pro/this-cheap-egpu-docking-station-uses-tb5-but-you-will-need-an-external-power-supply-to-get-it-working
- https://www.youtube.com/watch?v=feBWeSg0Nng
- https://www.ultrabookreview.com/34743-lenovo-legion-booststation/

Desktop AMD AM5

Desktop CPU

- https://www.anandtech.com/show/21524/the-amd-ryzen-9-9950x-and-ryzen-9-9900x-review/10
- https://www.amd.com/en/products/processors/desktops/ryzen/9000-series/amd-ryzen-9-9950x3d.html
- https://skatterbencher.com/2025/03/11/skatterbencher-85-ryzen-9-9950x3d-overclocked-to-5900-mhz/
- https://en.wikipedia.org/wiki/List_of_AMD_Ryzen_processors#Ryzen_AI_300_series
- https://en.wikipedia.org/wiki/Raptor_Lake
- https://www.youtube.com/watch?v=JbnBt_Aytd0

AM5 CPU cooler

AM5 Motherboard

- https://www.amd.com/en/products/processors/chipsets/am5.html
- https://pangoly.com/en/browse/motherboard?extra=Socket%3AAM5;FormFactor%3AATX;Chipset%3AAMD%20X870E;M2Ports%3A2,8&sort=low
- https://pcpartpicker.com/products/motherboard/#s=41&mt=ddr5&sort=-rammax&E=2,7&m=7,8,46&f=2&c=167

Workstation AMD sTR5/SP6

Workstation CPU

- https://www.amd.com/en/products/processors/workstations/ryzen-threadripper.html

sTR5/SP6 CPU cooler

- https://noctua.at/en/products/fan/nf-a12x25-ls-pwm
- https://www.thermaltake.com/aw420-aio-liquid-cooler.html
- https://thermaltakeusa.com/products/aw360-aio-liquid-cooler-cl-w450-pl12bl-a
- https://www.arukereso.hu/szamitogep-huto-c3094/noctua/nh-u14s-tr5-sp6-p1054940377/
- https://www.silverstonetek.com/en/product/info/coolers/xed120s_ws/
- https://noctua.at/en/nh-d9-tr5-sp6-4u/specification

TRX50/WRX90 Motherboard

- https://www.tomshardware.com/pc-components/ram/expansion-card-lets-you-insert-512gb-of-extra-ddr5-memory-into-your-pcie-slot-cxl-2-0-aic-designed-for-trx50-and-w790-workstation-motherboards
- https://www.pugetsystems.com/labs/articles/amd-trx50-vs-wrx90/
- https://www.corsair.com/us/en/explorer/gamer/gaming-pcs/amd-trx50-and-wrx90-motherboards-whats-the-difference/
- https://www.asus.com/motherboards-components/motherboards/workstation/pro-ws-trx50-sage-wifi/techspec/

TRX50/WRX90 Memory

- https://www.kingston.com/en/memory/search/model/109774/gigabyte-trx50-ai-top-motherboard?status_2=active
- https://www.kingston.com/en/memory/search/model/109774/gigabyte-trx50-ai-top-motherboard?status_2=active&capacity_2=64
- https://bizon-tech.com/bizon-x5500.html
- https://bizon-tech.com/gpu-benchmarks/NVIDIA-RTX-4090-vs-NVIDIA-RTX-4500-Ada/637vs697

Server AMD SP5

SP5 CPU

- https://www.amd.com/en/products/specifications/server-processor.html
- https://forum.level1techs.com/t/amd-epyc-9005-series-sp5-single-slot-or-dual-slot-motherboard-suggestion/229954
- https://www.reddit.com/r/LocalLLaMA/comments/1fcy8x6/memory_bandwidth_values_stream_triad_benchmark/
- https://www.reddit.com/r/LocalLLaMA/comments/1h3doy8/stream_triad_memory_bandwidth_benchmark_values/
- https://www.reddit.com/r/LocalLLaMA/comments/1k57b1o/comment/moftfs0/?context=3
- https://www.senetic.hu/category/amd-cpu-epyc-9004-11151/?cat=amd_cpu_epyc_9004_tray_64848&f_page=1&f_size=48&f_order=price_asc
- https://www.senetic.hu/product/100-000000799
- https://www.senetic.hu/product/100-000000790

- https://www.supermicro.org.cn/en/support/resources/cpu-amd-epyc-9005-9004-7003
- https://www.arukereso.hu/processzor-c3139/amd/epyc-siena-48-core-2-0ghz-sp6-tray-100-000001174-p1019326387/#termek-leiras
- https://www.arukereso.hu/processzor-c3139/amd/epyc-siena-64-core-2-0ghz-sp6-tray-100-000001172-p1019325790/#termek-leiras
- https://www.arukereso.hu/processzor-c3139/f:amd-socket-sp6,amd-epyc/?orderby=1
- https://www.arukereso.hu/processzor-c3139/amd/epyc-siena-64-core-2-0ghz-sp6-tray-100-000001172-p1019325790/
- https://smicro.hu/amd-epyc-genoa-9554-dp-up-64c-128t-3-1g-256mb-360w-sp5-100-000000790-5
- https://smicro.hu/amd-epyc-genoa-9534-dp-up-64c-128t-2-45g-256mb-280w-sp5-100-000000799-5
- https://smicro.hu/amd-epyc-genoa-9554-dp-up-64c-128t-3-1g-256mb-360w-sp5-100-000000790-5
- https://smicro.hu/epyc-genoa-9004-5
- https://en.wikipedia.org/wiki/Epyc#Fifth_generation_Epyc_(Turin_and_Turin_Dense)
- https://a16z.com/building-an-efficient-gpu-server-with-nvidia-geforce-rtx-4090s-5090s/

SP5 CPU cooler

- https://www.silverstonetek.com/en/product/info/coolers/xe04_sp5/
- https://forums.servethehome.com/index.php?threads/cooler-recommendations-for-400w-sp5.43530/
- https://smicro.hu/supermicro-snk-p0084ap4
- https://www.coolserver.com.cn/en/product_view_397_283.html
- https://www.arctic.de/en/Freezer-4U-SP5/ACFRE00158A
- https://www.phoronix.com/review/arctic-freezer-4u-sp5
- https://www.aliexpress.com/item/1005006621774992.html
- https://www.coolserver.com.cn/en/product_view_598_283.html
- https://www.silverstonetek.com/en/product/info/coolers/xe360_sp5/
- https://www.silverstonetek.com/en/product/info/coolers/xed120s_ws/

SP5 Motherboard

- https://www.asrockrack.com/general/productdetail.asp?Model=TURIN2D16-2T#Specifications
- https://www.arukereso.hu/alaplap-c3128/asrock/sienad8-2l2t-p1050062953/
- https://www.gigabyte.com/Enterprise/Server-Motherboard/ME03-CE0-rev-10
- https://www.storagereview.com/news/asrock-rack-server-motherboards-with-amd-epyc-8004-series-processors-support-announced
- https://www.supermicro.com/en/products/motherboard/H13SSL-NT
- https://www.supermicro.com/en/products/motherboard/H13SSL-N
- https://www.asrockrack.com/general/productdetail.asp?Model=SIENAD8-2L2T#Specifications
- https://www.gigabyte.com/Enterprise/Server-Motherboard/ME03-CE1-rev-10
- https://www.reddit.com/r/homelab/comments/1aeio2n/amd_epyc_siena_cpus_6_memory_channels_and_asrock/
- https://www.senetic.hu/product/SIENAD8-2L2T
- https://www.asrockrack.com/general/productdetail.pl.asp?Model=GENOA2D24G-2L%2B#Specifications
- https://www.reddit.com/r/LocalLLaMA/comments/1e5g65f/i_found_a_nice_motherboard_for_an_imaginary_gpu/
- https://www.supermicro.com/en/support/resources/cpu-amd-epyc-9005-9004-7003
- https://www.supermicro.com/en/products/motherboard/h13ssl-nt
- https://www.supermicro.com/Aplus/support/resources/OS/OS_Comp_H14_EPYC_9005_SP5-UP-1.cfm
- https://www.phoronix.com/review/supermicro-h13ssln-epyc-turin
- https://servers.asus.com/products/servers/server-motherboards/K14PA-U12
- https://www.gigabyte.com/Enterprise/Server-Motherboard/MZ73-LM0-rev-2x

- https://www.gigabyte.com/Enterprise/Server-Motherboard/MZ33-CP1-rev-3x
- https://www.reddit.com/r/HPC/comments/1bv0glb/epyc_genoa_memory_bandwidth_optimizations/
- https://youtu.be/bOxAdRfTpJg?t=500

SP5 Memory

- https://www.kingston.com/en/memory/search/model/106648/supermicro-h13ssl-nt-motherboard?status=active&capacity=32
- https://servertronic.de/index.php?&index=6&subindex=K0503&sockel=1&sys=200&gen=2
- https://servers.asus.com/products/servers/gpu-servers/ESC8000A-E12P
- https://servers.asus.com/products/servers/gpu-servers/ESC4000A-E12
- https://servers.asus.com/products/servers/gpu-servers/ESC8000A-E12
- https://www.reddit.com/r/homelab/comments/1h1iprj/epyc_97x4_genoa_motherboard/
- https://bizon-tech.com/bizon-x8000.html#specs
- https://store.supermicro.com/us_en/h13-2u-a-hyper-as-2025hs-tnr.html
- https://store.supermicro.com/us_en/h13-2u-a-hyper-as-2125hs-tnr.html
- https://www.exxactcorp.com/Exxact-TS2-106399742-E106399742
- https://www.tomshardware.com/pc-components/gpus/nvidia-unveils-2u-rtx-pro-6000-servers-at-siggraph-double-shot-of-blackwell-power-comes-to-smaller-rackmount-form-factors
- https://www.techpowerup.com/339818/nvidia-rtx-pro-servers-with-blackwell-coming-to-worlds-most-popular-enterprise-systems
- https://www.youtube.com/watch?v=HQ2EEQkbk8Y
- https://www.tomshardware.com/reviews/amd-4th-gen-epyc-genoa-9654-9554-and-9374f-review-96-cores-zen-4-and-5nm-disrupt-the-data-center/3
- https://www.youtube.com/watch?v=2Nq9r7qzwP8
- https://www.aime.info/en/shop/product/aime-g500-workstation/
- https://novatechgaming.com/products/novatech-rm600-dual-rtx-pro-6000-blackwell-192gb-vram-threadripper-pro
- https://www.falcon-nw.com/desktops/rak
- https://www.reddit.com/r/LocalLLaMA/comments/1mnevw3/pcimmiobar_resource_exhaustion_issues_with_2x_pro/
- https://www.reddit.com/r/LocalLLaMA/comments/1l6hnfg/4x_rtx_pro_6000_fail_to_boot_3x_is_ok/?share_id=6wS4_MYe-oljuQhs2MjYk&utm_name=ioscss

tinybox

- https://docs.tinygrad.org/tinybox/
- https://tinycorp.myshopify.com/products/tinybox-red
- https://www.reddit.com/r/LocalLLaMA/comments/1gcn3w9/a_glance_inside_the_tinybox_pro_8_x_rtx_4090/
- https://news.ycombinator.com/item?id=41365637
- https://archive.md/RTvI8
- https://www.tomshardware.com/tech-industry/artificial-intelligence/ai-accelerator-tinybox-pro-goes-up-for-preorder-for-usd40-000-the-device-features-eight-rtx-4090s-and-two-amd-genoa-epyc-processors
- https://www.tweaktown.com/news/97110/tinycorps-new-tinybox-ai-system-amd-gpu-starts-at-15k-nvidia-25k/index.html
- https://www.tomshardware.com/pc-components/gpus/tinybox-ai-accelerator-now-available-starting-at-dollar15k-available-in-7900xtx-and-rtx-4090-variants
- https://tinygrad.org/#tinybox
- https://www.techpowerup.com/news-tags/CDNA

Prices

- https://geizhals.eu/?m=1
- https://chatgpt.com/c/6830acbc-6258-800b-b8de-0b5a51a6ee13
- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=The%20biggest%20factor%20limiting%20performance,better%20suited%20for%20CPU%20inference
- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=In%20contrast%2C%20CPU%20inference%20uses,model%20on%20a%20capable%20GPU
- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=GPU%20inference%2C%20RAM%E2%80%99s%20primary%20use,RAM%E2%80%99s%20job%20is%20essentially%20done
- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=The%20biggest%20factor%20limiting%20performance,Threadripper%20PRO%20or%20EPYCs%2C%20are
- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=Regardless%20of%20the%20inference%20method%2C,be%20much%20of%20a%20concern
- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=In%20addition%20to%20strictly%20CPU,the%20greater%20the%20performance%20impact
- https://news.ycombinator.com/item?id=37067933#:~:text=Anything%20with%2064GB%20of%20memory,faster%20than%20you%20can%20read
- https://news.ycombinator.com/item?id=41046980#:~:text=match%20at%20L149%20A%20405B,disc%20and%20be%20unusably%20slow
- https://www.reddit.com/r/LocalLLaMA/comments/1cj4det/llama_3_70b_instruct_works_surprisingly_well_on/
- https://www.pugetsystems.com/labs/articles/tech-primer-what-hardware-do-you-need-to-run-a-local-llm/#:~:text=One%20of%20the%20first%20forks,be%20run%20via%20the%20GPU

PIM

- https://semiconductor.samsung.com/technologies/memory/pim/
- https://semiconductor.samsung.com/news-events/tech-blog/hbm-pim-cutting-edge-memory-technology-to-accelerate-next-generation-ai/
- https://www.techpowerup.com/278586/samsung-develops-industrys-first-high-bandwidth-memory-with-ai-processing-power
- https://www.techpowerup.com/328510/samsung-hopes-pim-memory-technology-can-replace-hbm-in-next-gen-ai-applications
- https://www.upmem.com/technology/
- https://www.upmem.com/
- https://sdk.upmem.com/master/00_ToolchainAtAGlance.html
- https://resources.sw.siemens.com/en-US/white-paper-innovative-upmem-pim-dram-requires-innovative-power-integrity-analysis/
- https://github.com/CMU-SAFARI/prim-benchmarks
- https://people.inf.ethz.ch/omutlu/pub/PrIM-UPMEM-Tutorial-Analysis-Benchmarking-SAFARI-Live-Seminar-2021-07-12-talk.pdf
- https://arxiv.org/html/2308.00846v3
- https://www.techpowerup.com/341384/french-team-develops-first-hybrid-memory-technology-enabling-on-chip-ai-learning-and-inference

Runners

Software

AMD CPU

- https://www.amd.com/en/developer/resources/technical-articles/zendnn-5-0-supercharge-ai-on-amd-epyc-server-cpus.html
- https://github.com/ggml-org/llama.cpp/issues/11744
- https://github.com/ggml-org/llama.cpp/discussions/11733

AMD GPU

- https://github.com/ROCm/flash-attention/tree/howiejay/navi_support
- https://github.com/ROCm/flash-attention
- https://llm-tracker.info/howto/AMD-GPUs
- https://www.youtube.com/watch?v=yCCoQ72DBpM

NVIDIA GPU

- https://developer.nvidia.com/nccl

Multi-GPU

- https://medium.com/@geronimo7/llms-multi-gpu-inference-with-accelerate-5a8333e4c5db
- https://docs.vllm.ai/en/v0.8.0/serving/distributed_serving.html

Library

- https://nn-512.com/
- https://github.com/explosion/thinc/tree/main
- https://github.com/ikawrakow/ik_llamafile
- https://github.com/ikawrakow/ik_llama.cpp/
- https://developer.nvidia.com/tensorrt

Framework

tinygrad

- https://tinygrad.org/#tinygrad
- https://news.ycombinator.com/item?id=33462337
- https://www.phoronix.com/forums/forum/linux-graphics-x-org-drivers/open-source-amd-linux/1519271-tiny-corp-nearing-completely-sovereign-compute-stack-for-amd-gpus-with-tinygrad/page2
- https://www.youtube.com/@geohotarchive/videos
- https://docs.tinygrad.org/showcase/
- https://github.com/tinygrad/tinygrad
- https://www.youtube.com/watch?v=Xtws3-Pk69o&list=PLzFUMGbVxlQsh0fFZ2QKOBY25lz04A3hi

TextSynth Server

- https://bellard.org/ts_server/
- https://bellard.org/ts_server/ts_server.html
- https://textsynth.com/documentation.html#embeddings
- https://textsynth.com/documentation.html#engines
- https://huggingface.co/fbellard/ts_server/tree/main
- https://github.com/sgl-project/sglang
- https://docs.sglang.ai/index.html
- https://dlib.net/ml.html#add_layer

- https://github.com/sony/nnabla
- https://www.nomic.ai/gpt4all
- https://www.phoronix.com/news/Red-Hat-llm-d-AI-LLM-Project
- https://github.com/unslothai/unsloth
- https://moondream.ai/
- https://llava-vl.github.io/
- https://github.com/bitsandbytes-foundation/bitsandbytes
- https://github.com/Mozilla-Ocho/llamafile
- https://justine.lol/oneliners/

Text

Images

- https://lightning.ai/pages/community/serve-stable-diffusion-three-times-faster/
- https://wan.video/

Voice

- https://huggingface.co/mistralai/Voxtral-Mini-3B-2507
- https://huggingface.co/mistralai/Voxtral-Mini-3B-2507#vllm-recommended
- https://mistral.ai/news/voxtral
- https://github.com/mozilla/DeepSpeech

OCR

- https://medium.com/@fuelyourdigital/how-to-extract-data-from-a-graph-image-with-chatgpt-4-6a26351e4227
- https://apps.automeris.io/wpd4/
- https://automeris.io/
- https://automeris.io/docs/
- https://zertrin.org/webplotdigitizer/
- https://github.com/automeris-io/WebPlotDigitizer
- https://web.eecs.utk.edu/~dcostine/personal/PowerDeviceLib/DigiTest/index.html
- https://www.colliseum.net/WebPlot/
- https://alternativeto.net/software/graphclick/?platform=windows
- https://alternativeto.net/software/graphclick/
- https://academia.stackexchange.com/questions/7671/software-for-extracting-data-from-a-graph-without-having-to-click-on-every-singl
- https://www.im2graph.co.il/
- https://github.com/Cvrane/ChartReader
- http://www.graphreader.com/
- https://plotdigitizer.com/
- https://www.microsoft.com/en-us/research/publication/chartocr-data-extraction-from-charts-images-via-a-deep-hybrid-framework/
- https://mathematica.stackexchange.com/questions/102362/reconstruct-a-graph-from-an-image-curved-edges-and-edge-crossings
- https://mathematica.stackexchange.com/questions/102262/how-to-generate-a-graph-from-a-picture-of-a-graph?lq=1
- https://www.geeksforgeeks.org/image-reconstruction-using-singular-value-decomposition-svd-in-python/
- https://medium.com/@pranjallk1995/pca-for-image-reconstruction-from-scratch-cf4a787c1e36

- http://www.fmwconcepts.com/imagemagick/textcleaner/index.php
- https://willus.com/blog.shtml?tesseract_accuracy
- https://www.zdnet.com/article/the-next-decade-in-ai-gary-marcus-four-steps-towards-robust-artificial-intelligence/
- https://github.com/zacharywhitley/awesome-ocr
- https://en.wikipedia.org/wiki/OCRopus
- https://en.wikipedia.org/wiki/Tesseract_(software)
- https://upmath.me/
- https://github.com/bytedance/Dolphin
- https://github.com/NanoNets/docext
- https://nanonets.com/research/nanonets-ocr-s/
- https://github.com/inferless/nanonets-ocr-s
- https://docs.inferless.com/introduction/introduction
- https://docstrange.nanonets.com/
- https://github.com/NanoNets/docstrange?tab=readme-ov-file
- https://nanonets.com/

RAG

- https://www.amd.com/en/developer/resources/technical-articles/2025/rag-with-hybrid-llm-on-amd-ryzen-ai-processors.html
- https://llm-tracker.info/RTX-PRO-6000

Embedding

- https://radimrehurek.com/gensim/auto_examples/core/run_core_concepts.html
- https://github.com/piskvorky/gensim
- https://radimrehurek.com/gensim/#
- https://en.wikipedia.org/wiki/ELMo
- https://en.wikipedia.org/wiki/Cosine_similarity
- https://en.wikipedia.org/wiki/Word2vec
- https://en.wikipedia.org/wiki/FastText
- https://en.wikipedia.org/wiki/Gensim
- https://en.wikipedia.org/wiki/Keras
- https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software

TinyML

- https://markaicode.com/no-llm-tinyml-models/

Papers

- https://arxiv.org/html/2505.07203v1
- https://github.com/d893a/localscore-leaderboard
- https://github.com/d893a/local-ml-models/blob/main/local_ml_hardware_alternatives.md#workstation
- https://geohot.github.io/blog/jekyll/update/2025/02/19/nobody-will-profit.html
- https://news.ycombinator.com/item?id=37067933#:~:text=Anything%20with%2064GB%20of%20memory,That%20should%20generate%20faster%20than