

IA local para PyMEs · Módulo 1

Elección de hardware

Los tres números que deciden todo, y cómo comprar —para uno mismo y para un cliente— sin casarse con una marca.

Qué se construye en este módulo

Los tres números: ancho de banda, capacidad y FLOPS

Estimar tokens/s y TTFT en papel

Por qué el pre-fill domina en cargas documentales y agénticas

Medir prefill y decode por separado

Comparación razonada de cinco topologías

Decidir entre GPU de consumo, memoria unificada CUDA, Apple Silicon, mini PC y SBC

Qué se construye en este módulo

 **Separación de
nodos y compra
para un cliente**

Aplicar la regla
cómputo $\neq$ servicios
 $\neq$ almacenamiento

1.1 · Los tres números: ancho de banda, capacidad y FLOPS

Los tres números: ancho de banda, capacidad y FLOPS

Cualquier equipo para inferencia local se describe con tres números. Con ellos se puede predecir en papel, con un margen de $\pm 30\%$, cómo va a rendir **antes de comprarlo**. Sin ellos se compra por fe.

En una tabla

Número	Unidad	Qué decide	Dónde s
Capacidad de memoria	GB	si el modelo + su con- texto caben	ficha té (GPU) o ficada
Ancho de banda de memoria	GB/s	la velocidad de deco- de (tokens/s)	ficha técn <i>bandwid</i>
Cómputo	TFLOPS (FP16/BF16)	la velocidad de prefill (TTFT)	ficha te <i>perform</i>

Cuándo NO usar esto

Los tres números: ancho de banda, capacidad y FLOPS

- No use la calculadora para comparar software (dos runtimes en el mismo equipo): eso se mide.
- No compre por TFLOPS si la carga es chat corto: ahí manda el ancho de banda.
- No confíe en la cifra de memoria «compartida» de un portátil: la GPU integrada usa RAM lenta.

Qué se degrada en cada perfil

A · sin GPU

Cabe lo pequeño; el prefill de documentos largos es inviable para uso diario

B · 8–16 GB VRAM

Modelos medianos cuantizados; contexto limitado por la caché; prefill aceptable

C · 24 GB+

Modelos grandes o medianos en Q8 con contexto amplio; prefill en segundos

1.2 · Por qué el prefill domina en cargas documentales y agénticas

Por qué el prefill domina en cargas documentales y agénticas

Casi todas las comparativas que circulan miden **tokens por segundo de salida**. Es el número equivocado para una PyME que quiere leer sus documentos. En cargas documentales y agénticas, lo que domina es el **prefill**: el tiempo de procesar la entrada antes de escribir la primera palabra.

En una tabla

Carga	Manda	Diseño
Chat corto (preguntas sueltas)	decode	ancho de banda equipo
Resumir, extraer, responder con RAG	prefill	cómputo; costo posible
Agentes de varios pasos	prefill × pasos	cómputo y costo

El comando, copiable

```
cd ~/labs/el-tornillo && mkdir -p labs/01-02 && cd labs/01-02
# -p = tokens de prefill a medir, -n = tokens de decode a medir
llama-bench -m <modelo>.gguf -p 512,2048,6144 -n 128 -o json > bench.json
python3 - <<'PY'
import json
for r in json.load(open("bench.json")):
    tipo = "prefill" if r["n_prompt"] > 0 else "decode "
    n = r["n_prompt"] or r["n_gen"]
    print(f'{tipo} {n:5d} tokens → {r["avg_ts"]:8.1f} tok/s  ({n/r["avg_ts"]:5.1f} s)')
PY
```

Si no corre desde cero con un comando, no entra al curso.

Cuándo NO usar esto

Por qué el prefill domina en cargas documentales y agénticas

- No optimice prefill si la carga real es chat corto: mida primero qué hacen los usuarios.
- No mida con un prompt de 50 tokens y extrapole: el prefill no es lineal en todos los runtimes.

Qué se degrada en cada perfil

A · sin GPU

Prefill de documentos largos en minutos: solo sirve para lotes nocturnos

B · 8–16 GB VRAM

Prefill en segundos; la caché de prefijos hace usable un agente de pocos pasos

C · 24 GB+

Prefill sub-segundo; agentes de muchos pasos y varios usuarios

1.3 · Comparación razonada de cinco topologías

Comparación razonada de cinco topologías

Las marcas cambian cada seis meses; las **clases** no. Cada una tiene un rol en la topología y un «cuándo no». Todo lo que sigue va con [VOLÁTIL – verificado: 2026-09] en cifras; el criterio es estable.

Cuándo NO usar esto

Comparación razonada de cinco topologías

- No diseñe todo en una caja. El nodo de cómputo se apaga para actualizar drivers; el proxy y la base de datos no pueden apagarse con él. Lección siguiente.
- No compre la clase 2 o 3 «porque tiene mucha memoria» para un caso de 5 usuarios con un 8B: una GPU de consumo de 16 GB rinde más por menos.
- No ponga un NAS a computar (módulo 2).

1.4 · Separación de nodos y compra para un cliente

Separación de nodos y compra para un cliente

Tres funciones, tres cajas (o al menos tres máquinas virtuales bien separadas, si el presupuesto obliga). Razones concretas:

En una tabla

Si se juntan...	Pasa esto
cómputo + servicios	actualizar el driver de la GPU tumba el proxy, el SSO y la base de datos: nadie entra
cómputo + almacenamiento	la GPU calienta los discos; un reinicio por un modelo colgado impide el acceso a los archivos
servicios + almacenamiento	aceptable en PyMEs pequeñas si el NAS tiene contenedores; el NAS no debe computar

El comando, copiable

```
cd ~/labs/el-tornillo && mkdir -p labs/01-04  
# ... escribir especificacion.md ...  
git add labs/01-04 && git commit -qm "Capa 1: especificación de hardware" && git log  
--oneline -1
```

Si no corre desde cero con un comando, no entra al curso.

Cuándo NO usar esto

Separación de nodos y compra para un cliente

- No separe en tres cajas físicas si el cliente tiene un empleado y un caso de uso: dos máquinas virtuales en un PC con GPU bastan; la regla de separación se cump
- No compre la GPU antes de tener trazas que digan cuánto se usa.

Qué se degrada en cada perfil



A · sin GPU



**B · 8–16 GB
VRAM**

3–5 usuarios



C · 24 GB+

10+ usuarios, modelos grandes

Entregables del módulo

- `calculadora.csv` con los tres equipos que usted tenga a mano o esté considerando.
- `bench.json` y las dos cifras de TTFT.
- la especificación, con el escalón 1+2 marcado como «fase 1» y el 3 como «fase 2».

**Los tres números que
deciden todo, y cómo
comprar —para uno mismo
y para un cliente— sin
casarse con una marca.**