MLOPSDAY BOLOGNA 2026 · COMMUNITY TALK

Il modello non ci sta: le leve per usarlo lo stesso

MoE più grandi della memoria: le idee comuni a tutti i motori

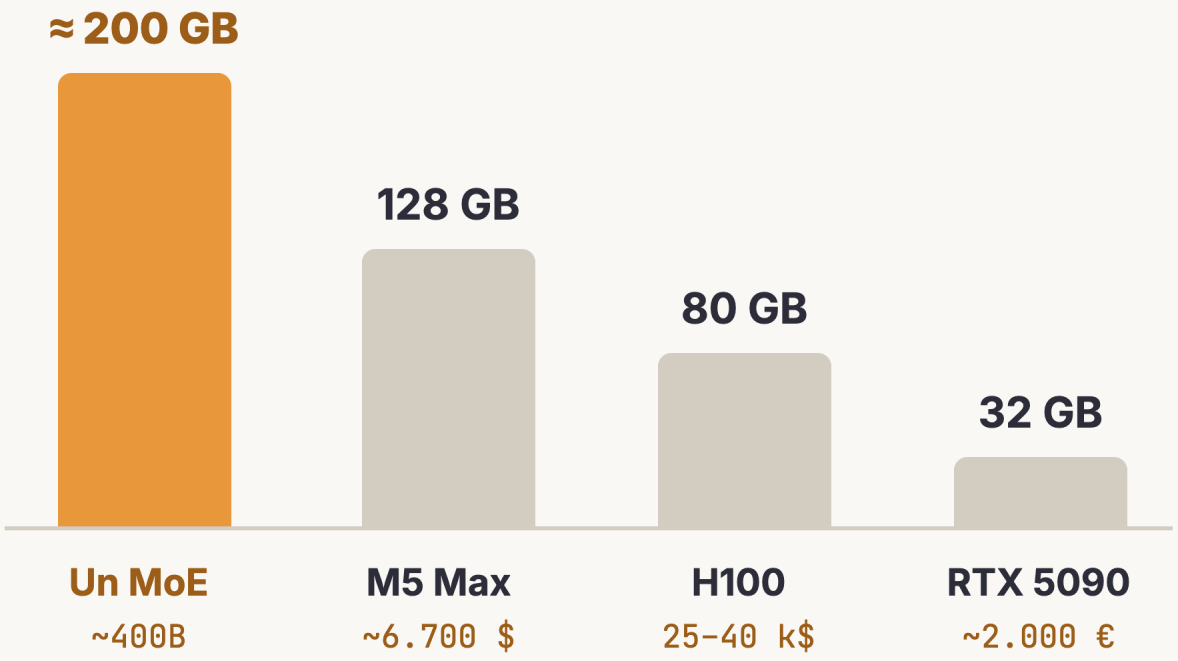
Francesco Montelli

Freelance Software Engineer · montelli.dev

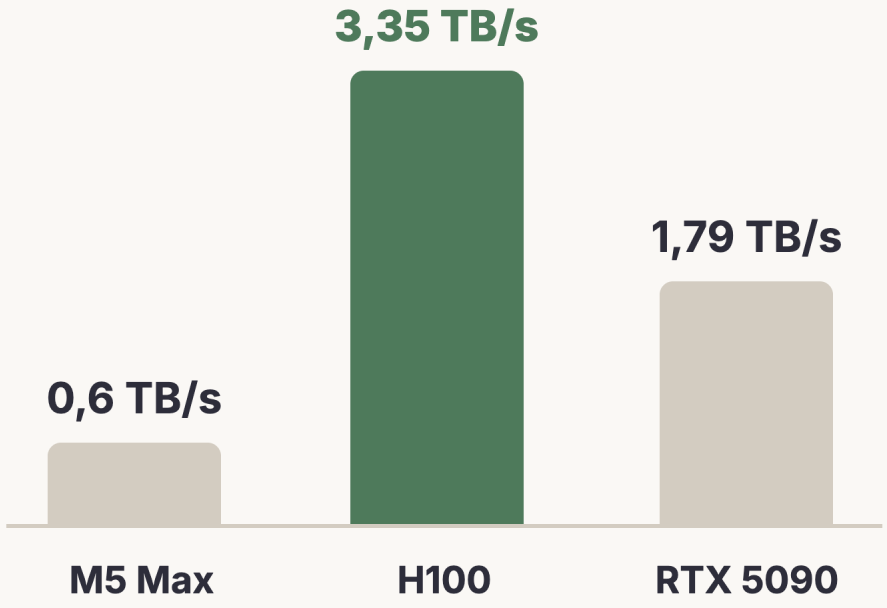
slide
e materiali

Il modello non sta in memoria

MEMORIA



BANDA



LUGLIO – OTTOBRE 2026

Un campo in rapida evoluzione

Flash-MoE

C · Metal

397B

su 48 GB

≈ 4,4 tok/s

MacBook M3 Max

Colibrì

C · solo CPU

744B

su 25 GB

0,05–1 tok/s

portatile, fino a M5 Max

llama.cpp

proposta, PR aperta

744B

su 128 GB

≈ 2 tok/s

GB10, memoria unificata

streamoe

MLX · Apple Silicon

35B

su 3–6 GB

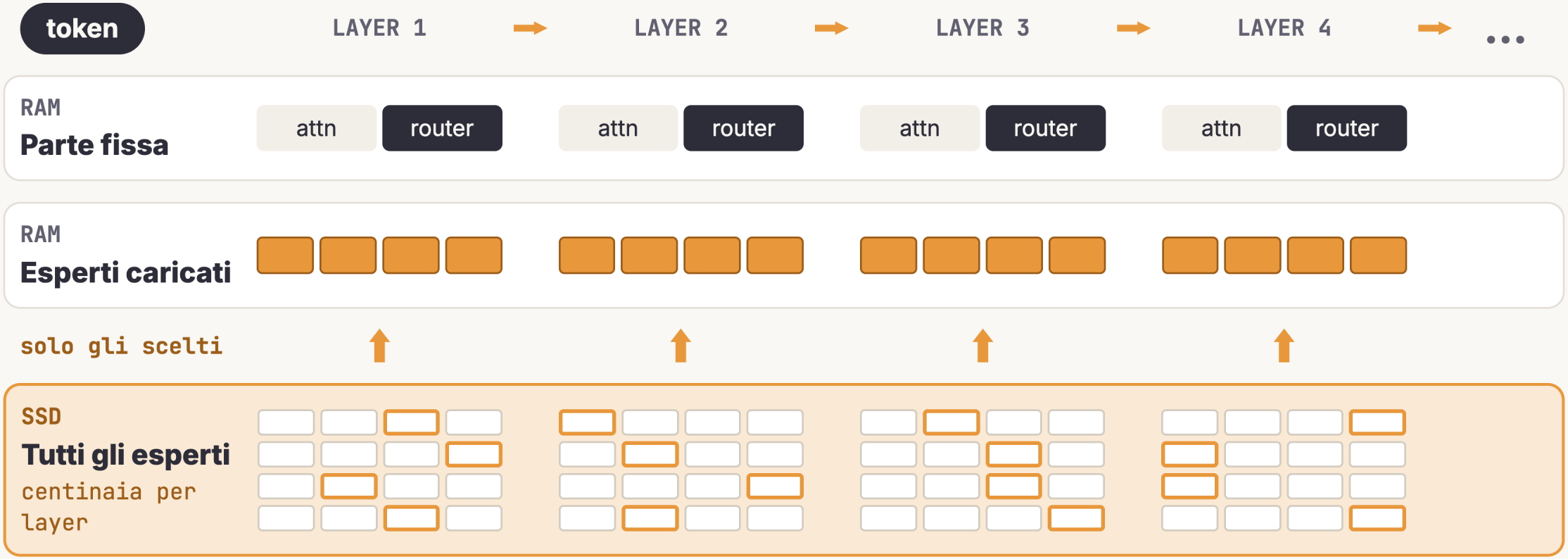
8–14 tok/s

MacBook M3 Pro

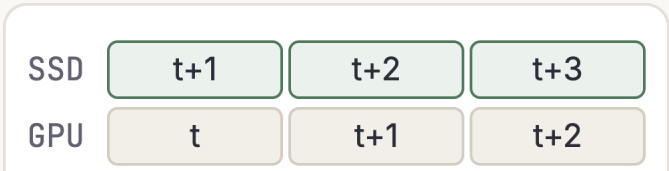
Meno bit, meno byte

BIT	PESO DI UN MOE DA ~400B	QUALITÀ
16	≈ 800 GB	originale
8	≈ 400 GB	quasi identica
4	≈ 200 GB	lieve perdita
2	≈ 100 GB	tool calling rotto

Fissa in RAM, esperti al bisogno



Caricare prima di sapere

STRATEGIA	COME	GUADAGNI	PAGHI
On-demand		**nessun errore**	**attesa a ogni layer**
Speculativo		**niente attesa, se indovini**	**se sbagli, due letture**
Prerouting		**lettura e calcolo insieme**	**il modello cambia un po'**

5 · RIEPILOGO

Tre leve, tre prezzi

LEVA	GUADAGNI	PAGHI
Quantizzazione	meno byte	sotto i 4 bit si rompe
Esperti su SSD	RAM ridotta	latenza
Caricare in anticipo	niente attesa	errori, o modello che cambia

Non serve più memoria.
Serve sapere cosa tenerci.

Parliamone
francesco@montelli.dev · montelli.dev
github.com/monte97

slide, paper e materiali