Open AI Infra
A Chapter of GCC

大语言模型系统级单位 token能耗测评

Global Computing Consortium
全球计算联盟

AI算力的另一面：从能力到能效

模型需求飞速增长

MMLU 从 60→90+，参数五年增长 5000 倍

中国日均token调用量：

2024 年初 1000 亿 →

2026 年 3 月 140 万亿

两年增长逾千倍

能耗正在成为天花板

2025 年全球数据中心用电激增 17%

推理已占机器学习计算 80-90% 需求

一个对话式 AI 年推理耗电量

可能超出训练数十倍

每百万 token 推理电费可达 1.4 元

占 API 定价的 7%-28%

未来竞争不仅是模型能力的竞争，更是能效与可持续的竞争

行业痛点：三重困境

选型困境

缺少方法论，选型靠经验

现状：选型靠经验

缺少方法论
决策不确定

大模型 × GPU = M × N 矩阵

大模型 (M)	GPU / 硬件 (N) 硬件1	硬件2	硬件3	硬件4	硬件5	.....
模型 1	✓	✓	✓	△	✗	....
模型 2	△	✓	△	✓	✗	
模型 3	✓	△	✗	✗	✗	
模型 4	✓	△	△	✗	✗	
模型 5	△	✗	△	✗	✗	
模型 6	✗	✗	△	✗	✗	
....	.	.	.	.	.	

✓ 适配较好　△ 勉强可用　✗ 不适配 / 风险高

缺乏系统化的匹配参考与量化评估体系
无法科学、客观地指导选型决策

榜单局限

只测能力，不测硬件

榜单聚焦模型能力，鲜有在相同硬件基线上横评实测

现有榜单只看能力分数

MMLU　GSMBK　HumanEval　....

模型 1	88.5
模型 2	86.2
模型 3	83.1
模型 4	81.4
...	...

缺失硬件实测表现

鲜有在相同硬件基线上
横评性能、时延、能耗的实测

性能 (吞吐/时延)　时延 (P95)　能耗 (Wh)

在相同硬件基线上横评

榜单只反映“能力上限”，不代表“部署效果”

场景复杂

场景负载不同，最优组合各异

不同场景对模型能力、上下文长度、生成长度、并发数、精度等要求差异显著

- 问答、摘要、写作等
- 短输入、低并发
- 延迟敏感度中等
- ……

- 代码生成、补全、调试等
- 长输入、中等并发
- 对上下文要求高
- ……

- 图像理解、生成、分析等
- 长（多模态）输入
- 高吞吐需求
- ……

场景不同，负载特征不同，最优组合各异

业界需要一个能同时回答**选什么模型、配什么硬件、跑什么场景**的决策系统

模型 × 硬件系统化匹配参考 + 相同硬件基线上的实测对比（性能 / 时延 / 能耗） + 面向场景的选型建议与最优组合

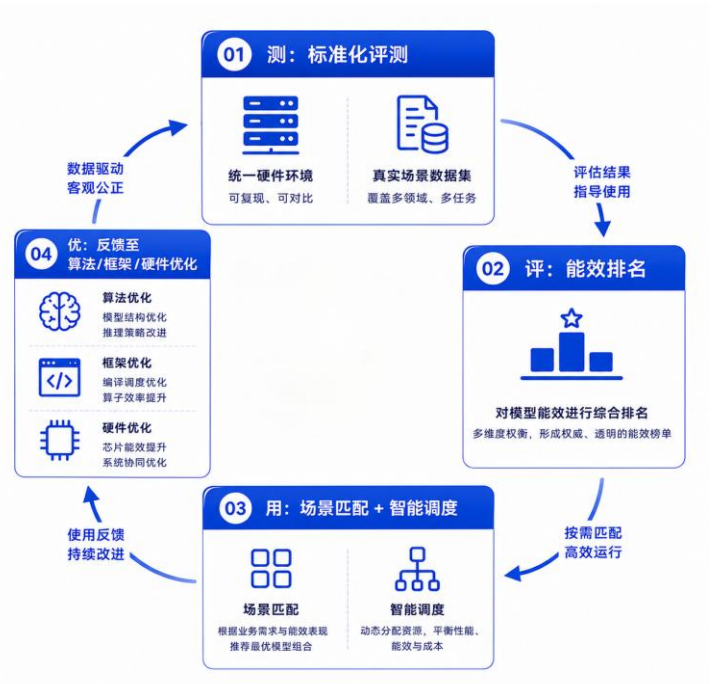
让选型更科学、决策更高效、部署更可靠、投入产出比更优

系统定位

差异化定位

维度	司南/HF 等榜单	本评测方案
硬件实测	✗ 未涵盖	✓ 实测
能耗评测	✗ 未涵盖	✓ 核心指标
B端选型	✗ 未涵盖	✓ 决策引擎
调度能力	✗ 无	✓ Aitra 调度

四层闭环结构



从“哪个模型能力最强”到“哪个组合最适合业务”

评测方法论：指标 · 场景 · 模型

五项核心指标

能耗
每token消耗焦耳量

GPU 利用率
硬件效率

TTFT
首 token 延迟

TPOT
解码速度

Throughput
系统吞吐

三类实际场景

W1 通用文本聊天
`lm-arena-chat`
`lmarena-ai/arena-human-preference-100k`
真实用户在 LMArena 上对 GPT-4 / Claude / Gemini 等模型的对话
256 请求
max_output = 1024 tokens

W3 代码补全 (FIM)
`sourcegraph-fim`
`sourcegraph/context-aware-fim-code-completions`
Sourcegraph 收集的真实代码库 FIM 补全任务，含丰富上下文
1024 请求
max_output = 2048 tokens

W4 多模态图像问答
`image-chat`
`lmarena-ai/VisionArena-Chat`
LMArena 视觉竞技场用户上传的真实图片 + 问题
1024 请求
max_output = 4096 tokens

八大主流模型

Gemma 4 31B
Dense + 视觉
量化：bf16
任务：文·代·图

Qwen 3.5 397B-A17B
MoE 384E
量化：FP8
任务：文·代

Llama 4 Maverick
MoE 17B×128E + 视觉
量化：FP8
任务：文·代·图

MiniMax-M2.7 REAP
MoE 10B 激活 + REAP 剪枝
量化：bf16
任务：文·代

DeepSeek-V4-Flash
MoE Hybrid CSA+HCA
量化：FP4 + FP8
任务：文·代

MiMo-V2-Flash
MoE Hybrid SWA+GA
量化：FP8 block
任务：文·代

Mistral-Medium-3.5-128B
Dense + Pixtral 视觉
量化：FP8 per-tensor
任务：文·代·图

gpt-oss-120b
MoE 5B 激活,128E + hybrid SWA
量化：MXFP4 4-bit
任务：文·代

评测方法论：环境 · 流程

统一测试环境

硬件
8 x NVIDIA

软件栈

- **ml-energy/benchmark v3.0**
  Benchmark 框架
- **vLLM 0.19.1 (Docker)**
  推理引擎
- **Zeus 0.13.1**
  能耗测量库
- **Prometheus + vLLM /metrics**
  运行时观测
- **transformers 5.5.4**
  Tokenizer / Config

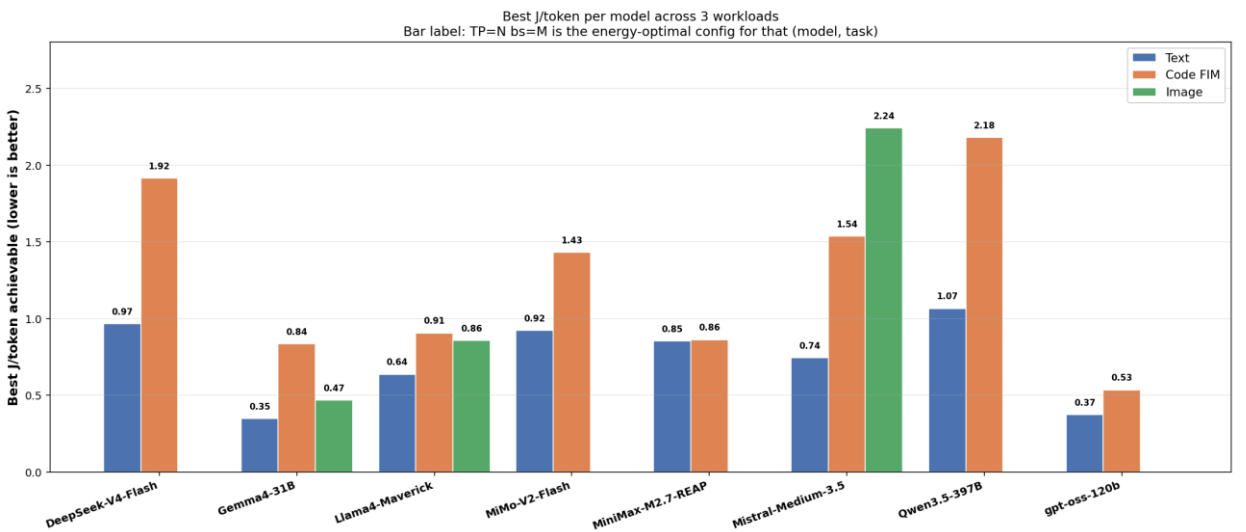
规范测试流程

1. 下载模型 + HF cache 准备
2. 启动 vLLM Docker 容器
3. 等待 /health 就绪 (≤ 600 s)
4. Zeus 开启双能耗窗口
5. 并发发送 256/1024 请求
6. Prometheus 旁路采样
7. 结束并写出产物
8. 回填元数据

三类场景 · 能耗总览

所有模型三场景最低 J/token（最优配置）

Model	参数量	精度	架构	文本	代码	多模态	特征标签
Gemma4-31B	31B	BF16	Dense	**0.35**	0.47	0.84	全局最低 · 跨任务稳定 · Dense标杆
gpt-oss-120B	120B	FP8	MoE	**0.37**	0.53	—	MoE+低比特 · 能效吞吐均衡
Llama4-Maverick	400B	FP8	MoE	0.64	0.91	0.86	多模态最佳 · 无短板
Qwen3.5-397B	397B	FP8	MoE	1.07	2.18	—	跨任务能耗高且波动剧烈
MiniMax-M2.7	172B	BF16	MoE	0.85	0.86	—	代码场景吞吐登顶 4300 tok/s
DeepSeek-V4	—	BF16	MoE	**0.97※**	**1.92**	—	※换SGLang(原vLLM 4.43) ↑5.7×吞吐
Mistral-Med-3.5	—	BF16	Dense	0.82	1.18	**2.28**	多模态能耗最高 · 视觉解码瓶颈
MiMo-V2-Flash	—	FP8	MoE	0.92	1.43	—	中等表现，框架适配待完善



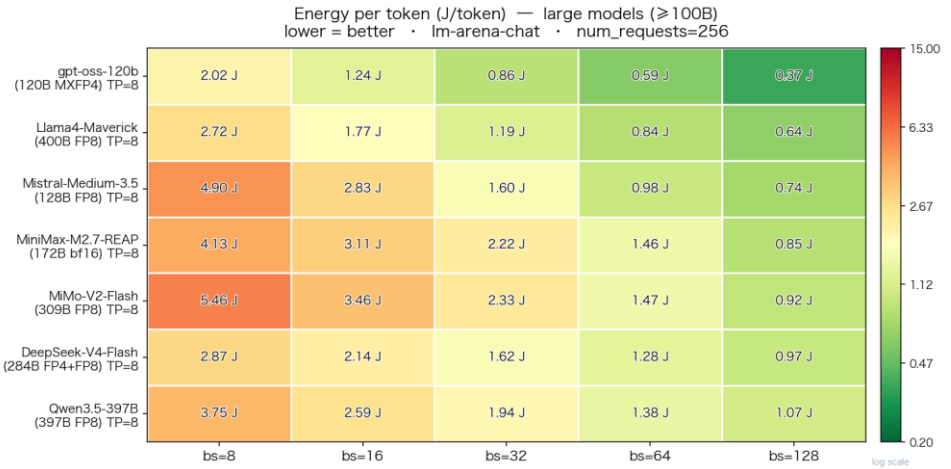
1 场景负载决定能耗基线

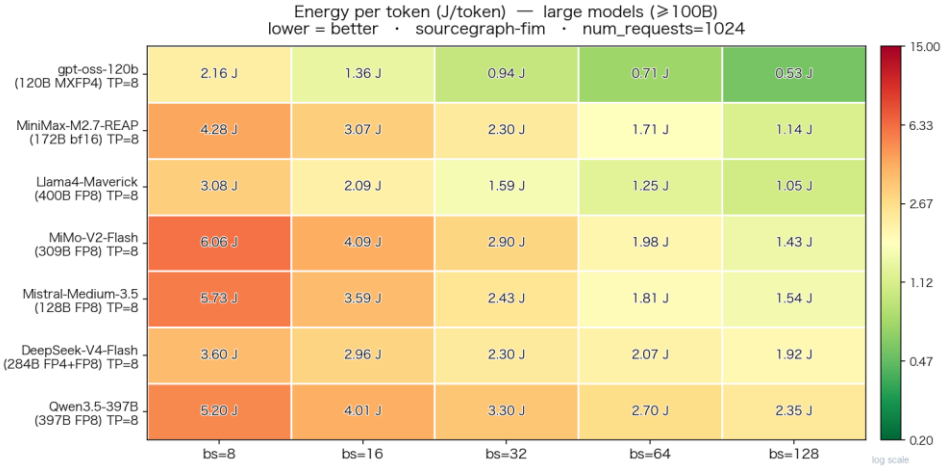
场景负载： 文本 < 图像 < 代码
能耗梯度： 文本 < 图像 < 代码

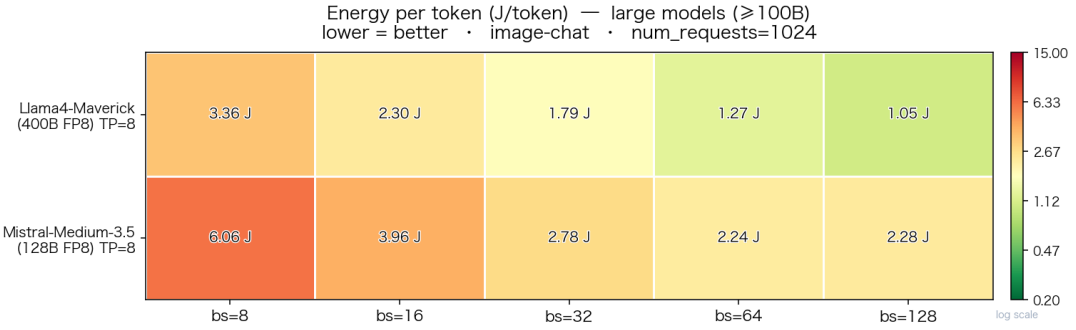
2 同一模型在不同场景下性能差异较大

跨任务稳定： Gemma + gpt + Llama
仅轻量场景优秀： Qwen + Mistral

热力图揭示的能耗规律







1 批处理大小是能效第一驱动力

能耗随批处理大小增大单调下降

合理配置并发数对降低推理成本的收益超过模型选型本身

2 同硬件下模型间能耗差距也可能较大

Qwen3.5-397B 与 Llama4-Maverick

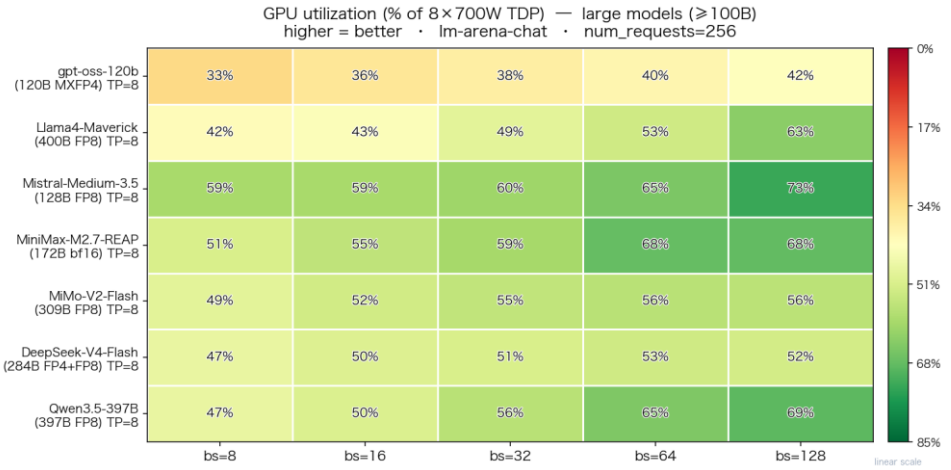
相同的硬件、相同的精度条件下，**能耗差距达到3倍以上**

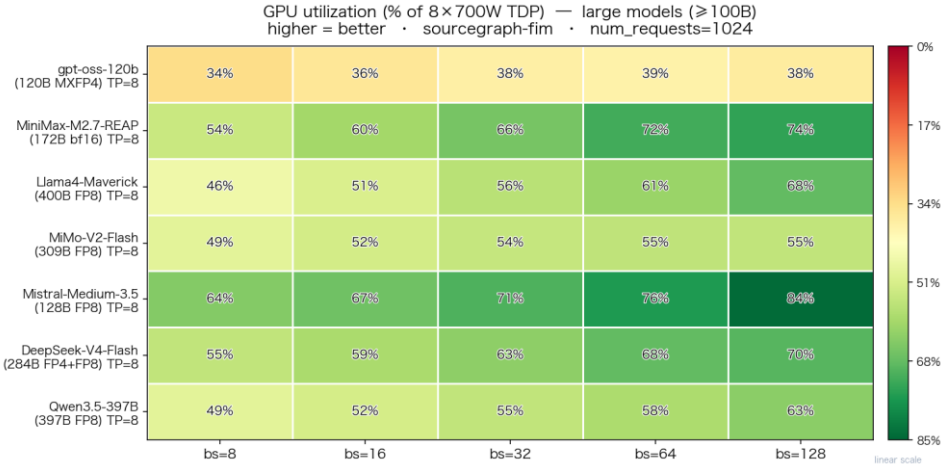
3 MoE + 低比特量化能效优势显著

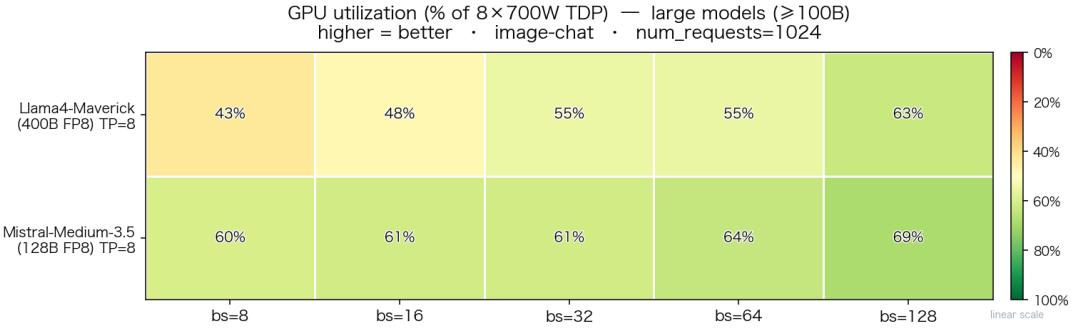
以 gpt-oss-120B 为代表的 MoE 配合 FP8 低精度路线

在能效和吞吐之间取得了最好的平衡

GPU 利用率与TPOT：能耗之外的权衡







1 批处理大小对GPU利用率影响有限

GPU利用率集中在 40-70%
批处理大小变化的影响有限

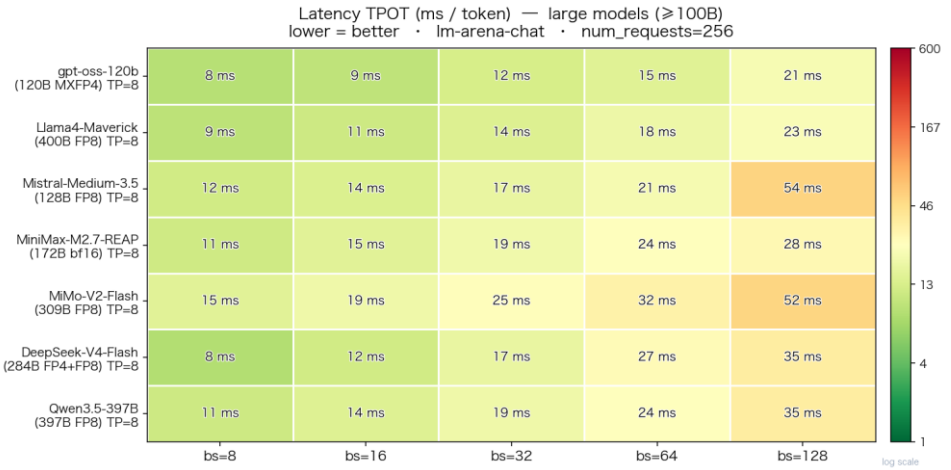
2 场景对GPU利用率影响较大

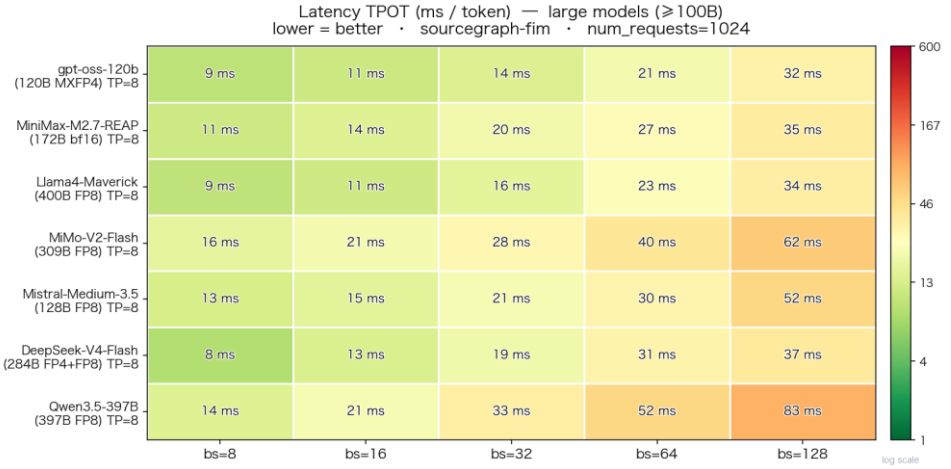
代码场景下GPU利用率整体高于文本场景
长提示词的密集预填充使GPU得到更充分的利用

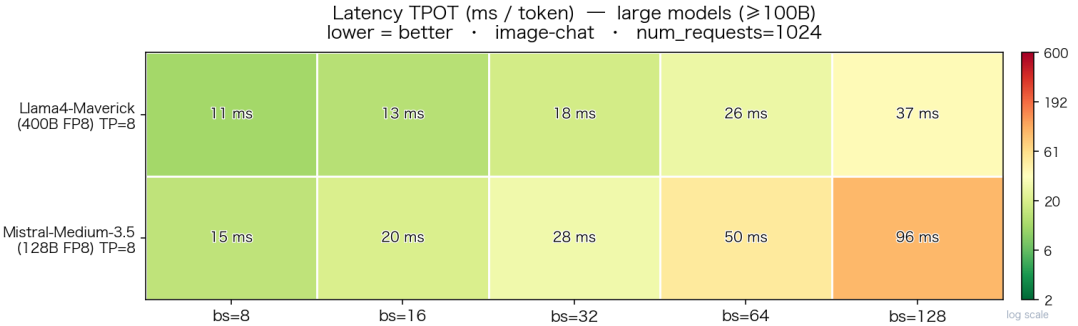
3 高GPU利用率 ≠ 高能效

gpt-oss-120b等GPU利用率低但能效好
Mistral-Medium-3.5等利用率高但能效差
能耗和GPU利用率之间存在一定的负相关性

GPU 利用率与TPOT：能耗之外的权衡







1 能耗-延迟的天然权衡

TPOT随批处理大小增大而增大，是**摊薄能耗的天然代价**
可能是由于高并发带来的计算资源争用

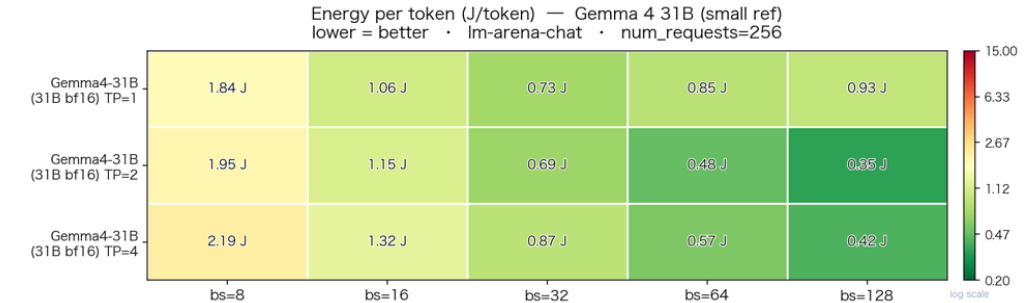
2 场景对大部分模型TPOT影响较小

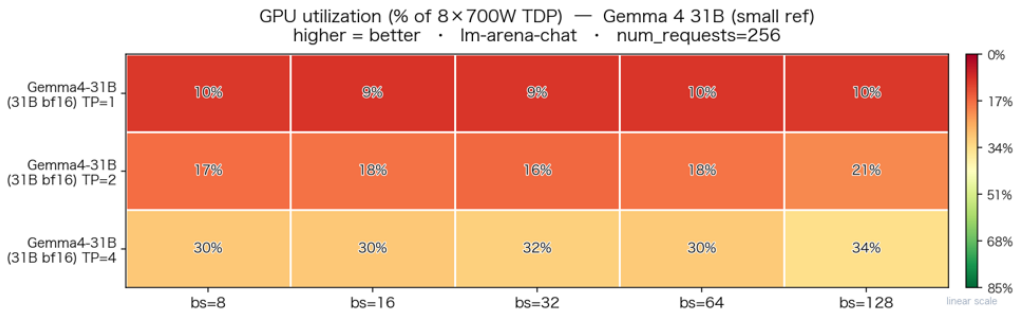
低并发时大多数模型在不同场景下的TPOT几乎一致
但部分模型在高并发时TPOT相较文本场景显著恶化

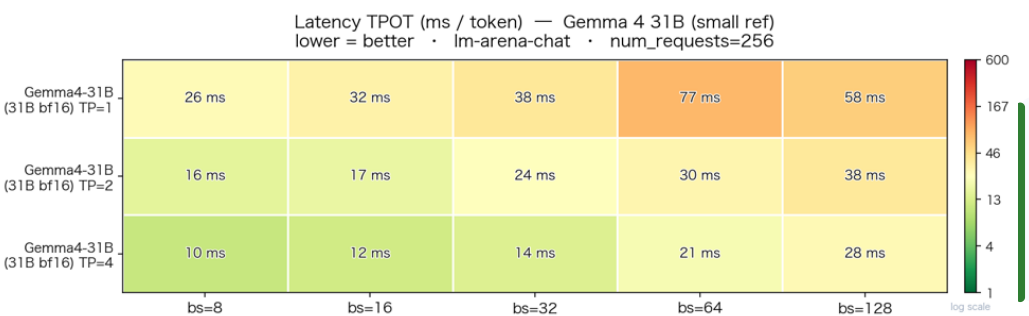
3 框架成熟度显著影响TPOT

框架适配不足的模型即使在低并发下TPOT也严重偏高
瓶颈可能在单请求解码效率而非并发处理能力

Gemma 深度分析：小模型的能效标杆







1 TP=2 是能效甜点

TP=4 因通信开销抵消了计算收益

TP=1 在大 BS 下出现能耗反弹，说明**单卡计算能力已经饱和**

2 过度分卡收益有限

Gemma4 只有 31B 参数，对 8 张芯片来说偏小

过度分卡不提升性能，反而因为通信开销拖累整体效率

3 BS增加延迟，TP抑制延迟

TP的增大会**显著降低TPOT**（BS=8时从26ms降至10ms）

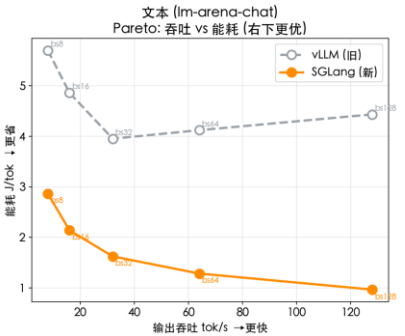
BS增大则会**推高TPOT**（TP=1时从26ms升至58ms）

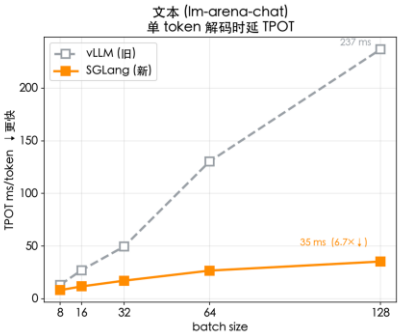
部署建议（文本聊天场景）

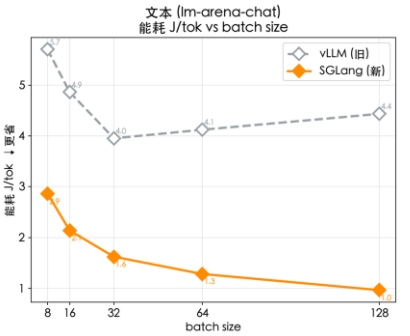
TP=2/4，BS=64/128

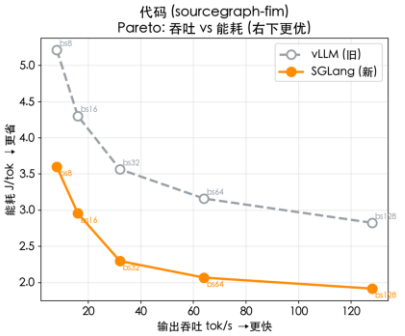
能耗、延迟、利用率三者最均衡的配置区间

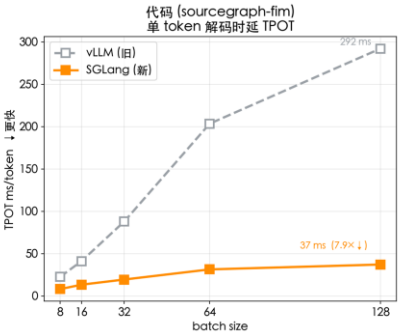
DeepSeek 深度分析：推理引擎对能效的影响

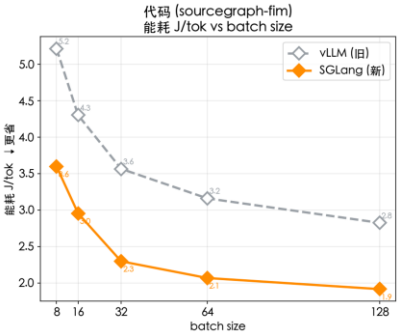












文本场景（BS=128）

指标	vLLM	SGLang	对比
吞吐(tok/s)	487	2763	↑5.7×
能耗(J/tok)	4.43	0.97	↓78%
TPOT(ms)	237	35	↓85%

代码场景（BS=128）

指标	vLLM	SGLang	对比
吞吐(tok/s)	764	1879	2.5×
能耗(J/tok)	2.83	1.92	↓32%
TPOT(ms)	292	37	↓87%

同一模型，换引擎 = 换表现

推理引擎对能效的影响可能超过模型架构本身

模型架构创新必须**与推理框架协同**，能效评测需区分**模型自身能效**和**当前框架适配下的实测能效**

从评测到落地：电费映射 + 生产选型

能效电费映射：从能效到可感知的电力成本

单位token电费 = 每token消耗电量 × 电价 × 数据中心电能利用效率

生产环境选型推荐

按业务场景挑选

业务场景	推荐模型	推荐配置	J/tok	理由
文本 · 能效极致	Gemma4-31B	TP=2 bs=128	**0.348**	全场最低能耗，2卡即可部署
文本 · 吞吐优先	gpt-oss-120B	TP=8 bs=128	0.373	4013 tok/s，能效吞吐均衡
代码 · 吞吐优先	MiniMax-REAP	TP=8 bs=256	0.86	4180 tok/s，代码吞吐登顶
代码 · 能效优先	gpt-oss-120B	TP=8 bs=128	0.535	3846 tok/s，代码场景能效最优
图像 · 能效优先	Gemma4-31B	TP=2 bs=128	**0.467**	多模态场景能效最优
图像 · 吞吐优先	Llama4-Maverick	TP=8	—	3365 tok/s，多模态吞吐第一

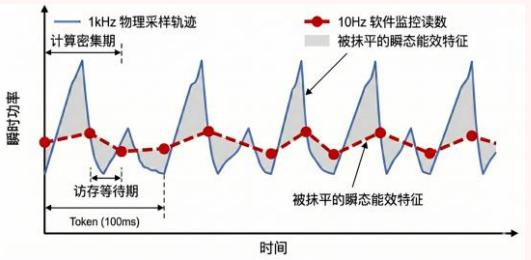
选型三原则

前瞻：Token级能耗精确定位

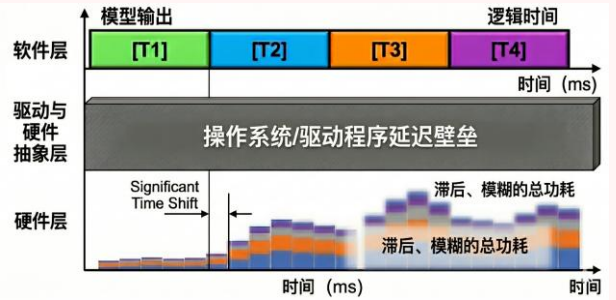
现有能耗监测的局限

当前能耗获取高度依赖 NVML 或类似系统驱动接口

1. 低频采样抹平能耗特征
   无法捕捉 <100ms 的 Token 级波动



2. 软件读数无法精准对应电流变化与计算任务细粒度算法调优缺乏反馈回路



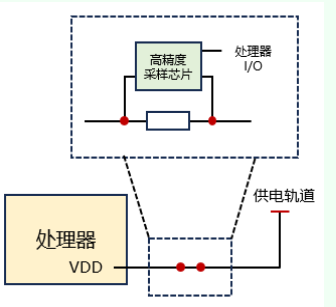
硬件方案：高精度能耗测量平台

摒弃传统 PMIC 读数，在芯片供电轨道上进行侵入式能耗测量

1. 供电轨串联精密分流电阻，利用高精度电流监控芯片进行 1kHz 实时采样
2. 采样数据经时间标定，实现 Token 级瞬态能量特征捕捉
3. RK3588 边缘计算节点平台完成概念验证

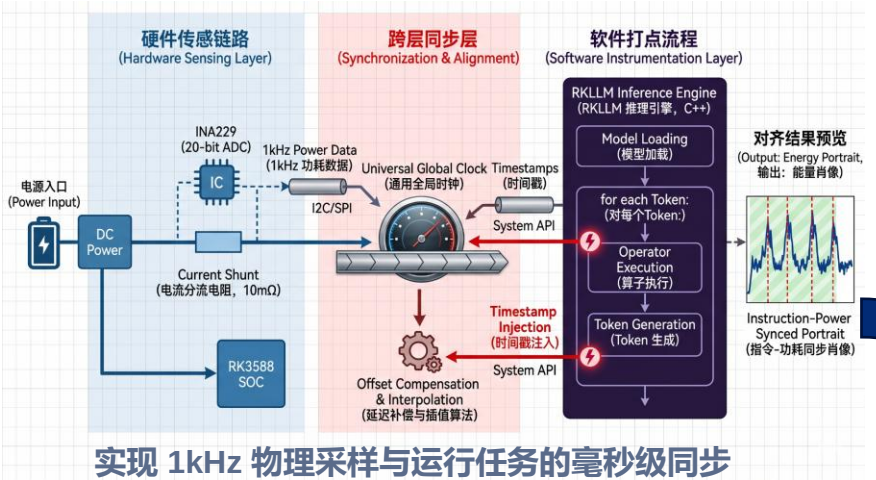
实物验证方案

硬件能耗测量探针方案

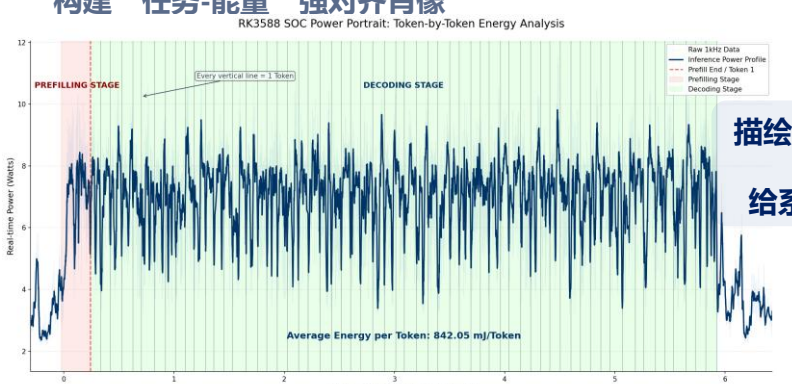


前瞻：Token级能耗精确定位

测量方法



实现 1kHz 物理采样与运行任务的毫秒级同步
彻底消除驱动延迟与 PMIC 滤波导致的能效表征模糊
构建“任务-能量”强对齐肖像



描绘模型运行不同阶段的能耗特征
定位Token能耗
给系统的能耗优化提供精确指导

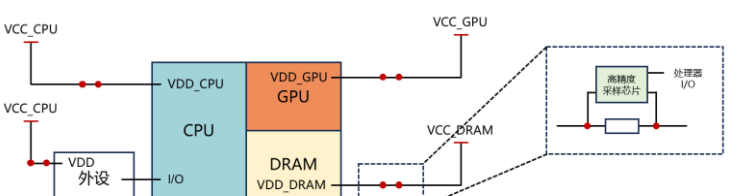
未来价值

1. **揭示大规模并行下的通信能耗墙**
   高频侵入式测量
   精确定位功耗峰值（Tensor Core矩阵运算/存储控制器总线切换）

2. **捕捉复杂调度下的长尾能耗效应**
   量化指令分配不均或缓存失效
   指导编译器算子重排

3. **支撑亚毫秒级动态电压频率调整**
   给计算阶段的预判提供数据支持，**支撑亚毫秒级的功率预测模型**
   在Token级计算负载到来前，**精准调整供电轨电压**

进阶能耗测量硬件方案



未来展望

局限

1 模型覆盖有限
现有评测仅覆盖主流模型，
新兴模型与长尾模型纳入不足。

2 硬件基线单一
硬件平台类型和规格有限，
难以反映真实多样化部署环境。

3 框架适配差异
不同推理框架对模型支持与优化能力不一，
影响评测公平性与可比性。

4 量化精度差异
不同量化方法和精度设置存在差异，
影响模型性能的可比性与稳定性。

5 测试负载代表性
测试集规模与场景覆盖有限，
难以全面反映真实业务负载特征。

未来

1 评测维度拓展
训练-微调-推理全生命周期
覆盖模型从训练、微调到推理的全流程，
构建更完整的评测体系。

2 硬件平台扩展
NPU/TPU/国产AI加速器
纳入更多异构计算平台，
提升评测的广度与工程参考价值。

3 标准与政策衔接
行业→国家→国际标准
推动评测标准与政策体系协同，
促进标准化、合规化与国际化。

4 “测-评-用-优”闭环构建
打通评测、评估、应用与优化全链路，
实现数据驱动的持续改进与价值闭环。

可量化的选型依据
为企业提供数据驱动的
模型×硬件匹配决策

绿色算力调度决策
为智算中心提供
能效驱动的调度支撑

推动能效系统优化
从能力单点突破
走向能效系统优化

致谢

•本次测评由**清华大学电机工程与应用电子技术系信息能源实验室、清华四川能源互联网研究院信息能源与绿色智算技术研究所、新型电力系统运行与控制全国重点实验室电化学储能高效集成与控制团队**联合牵头编制，**Linux基金会研究部门**和**SODA基金会**提供开放协作生态系统支持。

•在此谨向Linux基金会、SODA基金会致以诚挚感谢。

清華大學
清华四川能源互联网研究院
Sichuan Energy Internet Research Institute
Tsinghua University

Open AI Infra

Thank you!

Global Computing Consortium
全球计算联盟