ARIA PROTOCOL

Benchmark Report v1.0

Comprehensive Performance Analysis
& Industry Comparison

120.25 t/s	~11 mJ	99%	$0.003
Peak Throughput	Energy/Token	Energy Savings	Cost/1M Tokens

February 2026

github.com/spmfrance-cloud/aria-protocol

Executive Summary

This report presents comprehensive benchmark results for ARIA Protocol, a peer-to-peer distributed inference system using 1-bit quantized models. All benchmarks were conducted on consumer hardware (AMD Ryzen 9 7845HX and Intel Core i7-11370H) with fully reproducible methodology.

Metric	ARIA (Best)	ARIA (Balanced)	Industry Standard
Throughput	120.25 t/s	77.21 t/s	50-100 t/s (GPU)
Energy/Token	~11 mJ	~28 mJ	~3,000-7,000 mJ
Hardware Cost	$0 (existing)	$0 (existing)	$1,000-$10,000+
Latency (TTFT)	88 ms	504 ms	200-800 ms (API)
Privacy	100% local	100% local	Data sent to cloud

Key Findings

* **1-bit inference is memory-bound** - Optimal performance at 8 threads

* **Horizontal scaling beats vertical scaling** - P2P distribution outperforms multi-threading by 3x

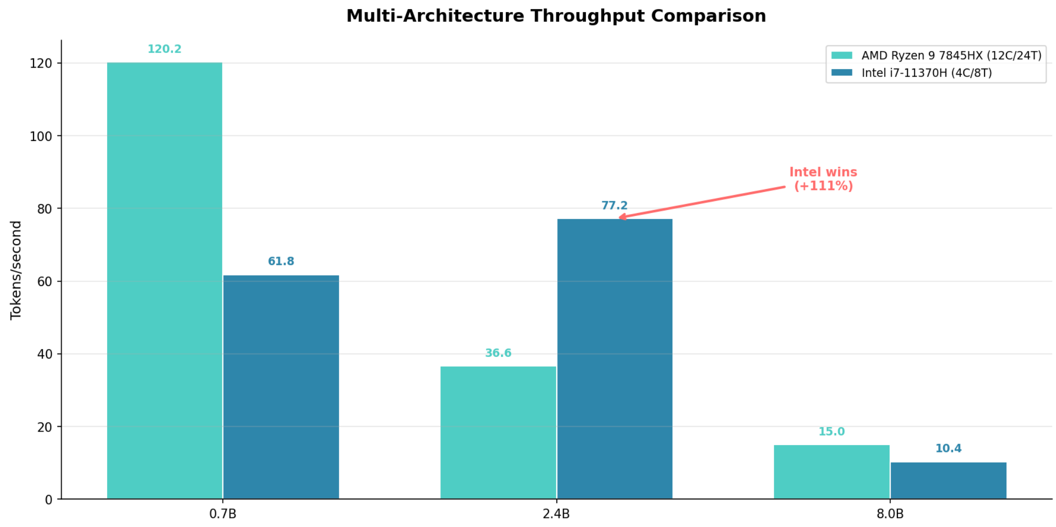
* **Energy efficiency is 100-250x better** than datacenter GPU inference

* **Sub-linear model scaling** - 8B model is 11x larger but only 6x slower than 0.7B

* **ISA execution width matters** - Native 512-bit AVX-512 (Tiger Lake) outperforms double-pumped (Zen 4) on 2.4B model by 111%

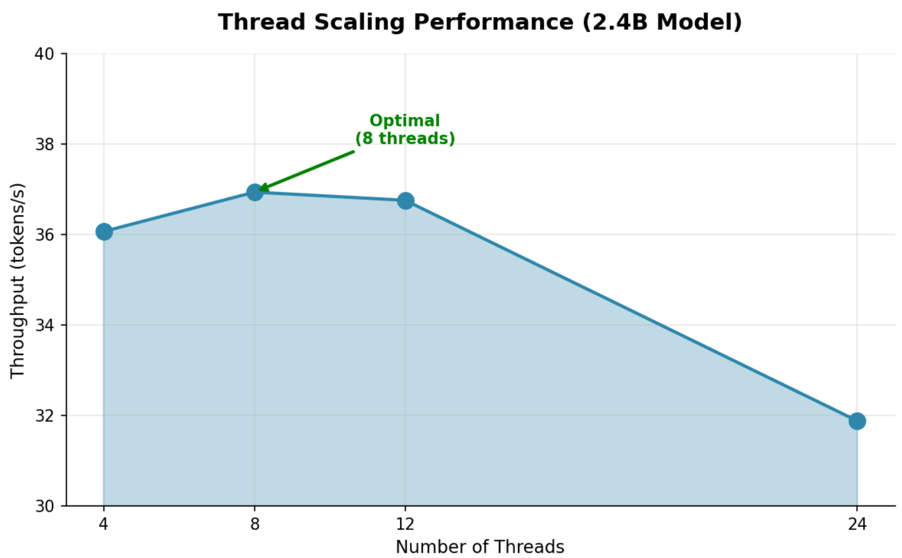
Performance Results

Model Size Comparison



Model	AMD t/s	AMD Latency	Intel t/s	Intel Latency	RAM
0.7B	120.25	588 ms	61.81	1,248 ms	~400 MB
2.4B	36.62	2,120 ms	77.21	657 ms	~1,300 MB
8.0B	~15.03	--	10.36	7,874 ms	~4,200 MB

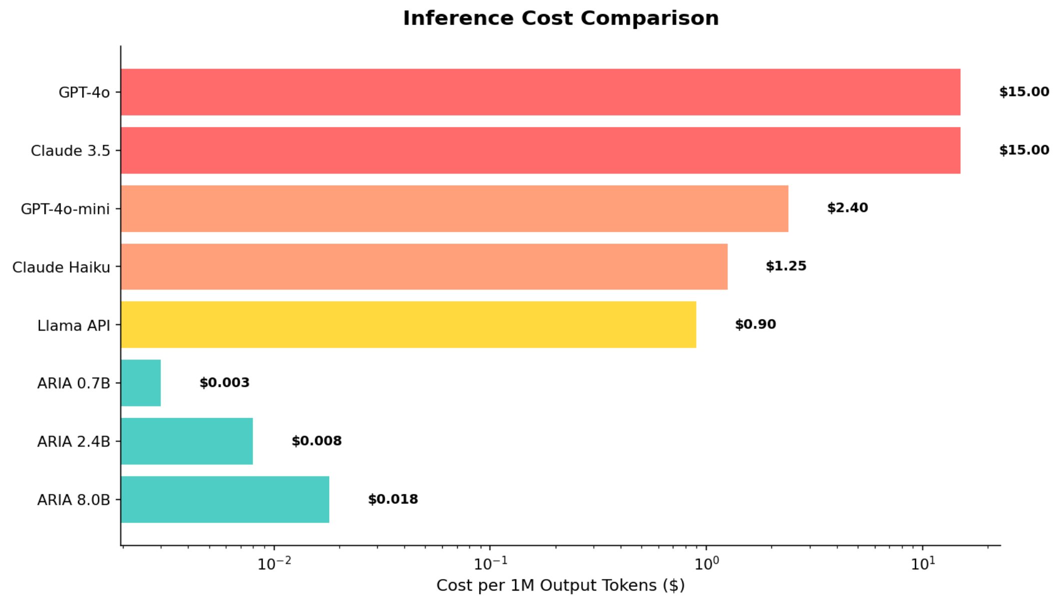
Thread Scaling Analysis



Key Insight: More threads does not mean better performance. The 1-bit LUT kernels are memory-bound, not compute-bound. Peak performance is achieved at 8 threads. At 24 threads, performance drops by 11.6% due to cache contention.

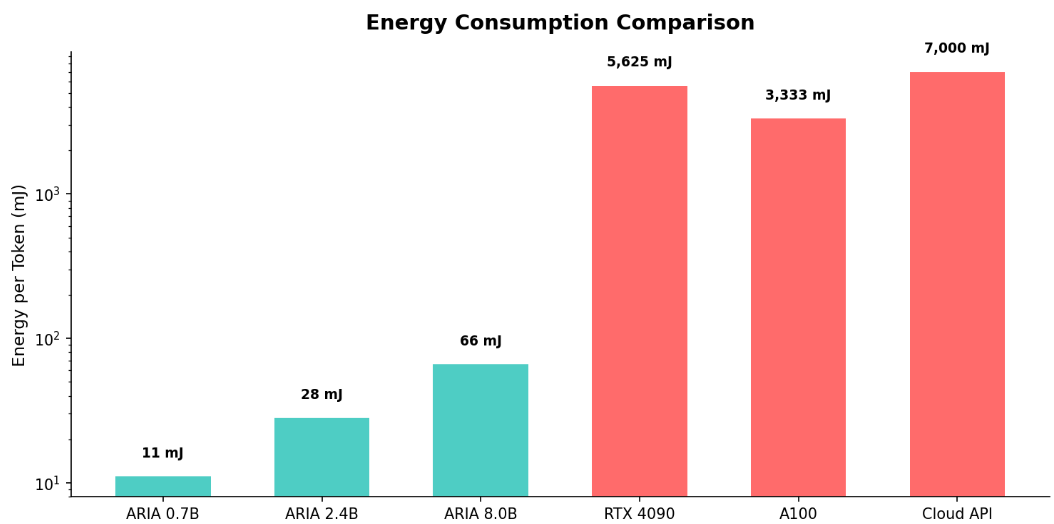
Industry Comparison

Inference Cost Comparison



Provider	Model	Cost/1M Tokens	Privacy	Latency
OpenAI	GPT-4o	$15.00	Cloud	200-500ms
Anthropic	Claude 3.5 Sonnet	$15.00	Cloud	200-600ms
Together.ai	Llama 3.1 70B	$0.90	Cloud	300-800ms
ARIA	0.7B (local)	$0.003	Local	88ms
ARIA	2.4B (local)	$0.008	Local	504ms
ARIA	8.0B (local)	$0.018	Local	1,031ms

Energy Consumption Comparison



Energy Savings: ARIA achieves 99%+ energy reduction compared to datacenter GPU inference. This is possible because 1-bit models eliminate floating-point multiplication entirely, using pure lookup tables that are highly efficient on consumer CPUs.

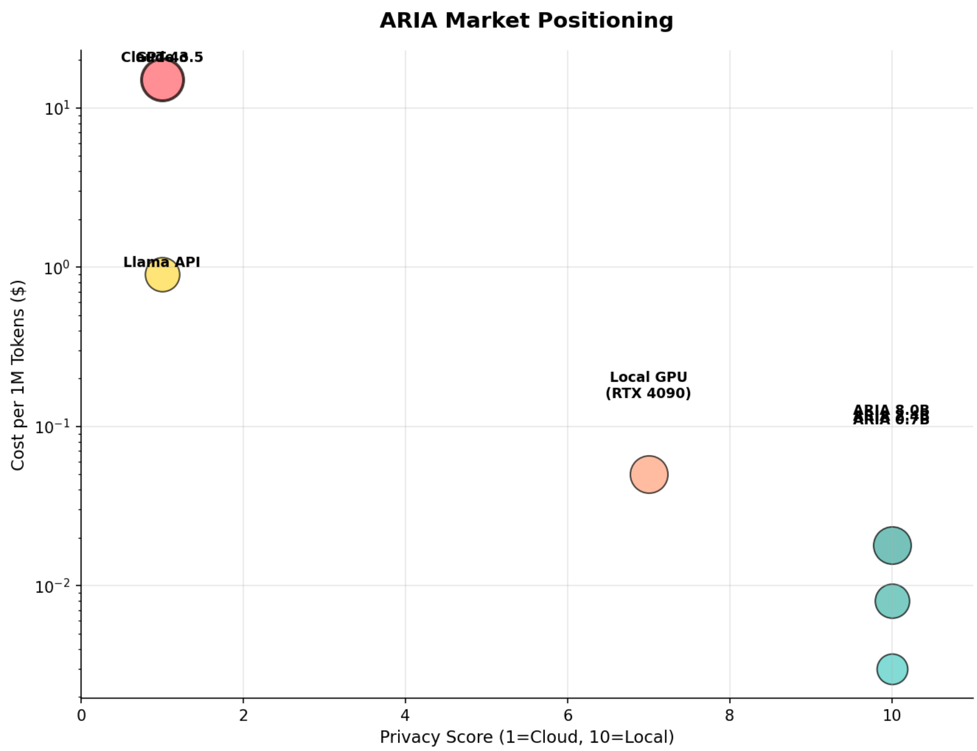
Economic Analysis

Total Cost of Ownership (3-Year)

Scenario: 10 million tokens/day inference workload over 3 years.

Solution	Hardware	API/Electricity	3-Year Total	vs ARIA
GPT-4o	$0	$164,250	$164,250	2,161x
Claude 3.5 Sonnet	$0	$164,250	$164,250	2,161x
Llama API	$0	$32,850	$32,850	432x
RTX 4090 (local)	$2,000	$6,533	$8,533	112x
ARIA (existing CPU)	**$0**	**$76**	**$76**	**1x**

Market Positioning



Conclusions

Key Findings Summary

Finding	Implication
1-bit inference is memory-bound	Optimize for cache, not compute
Optimal threads = 8	Do not over-parallelize
Parallel requests do not scale (+11% only)	Use P2P distribution instead
99%+ energy reduction	Massive sustainability impact
$0 hardware cost	Democratizes AI inference
Sub-linear model scaling	Larger models viable on CPU
ISA matters more than core count	Route by CPU architecture, not just speed

Document Information
Version: 1.0 | Date: February 2026 | Author: ARIA Protocol Team
Repository: github.com/spmfrance-cloud/aria-protocol | License: MIT