Harmonic Stack: Parallel Inference Scaling on Consumer Hardware

Ghost in the Machine Labs

All Watched Over By Machines Of Loving Grace

January 31, 2026

Abstract

We present benchmark results comparing parallel inference scaling on two consumer-grade AI platforms: the NVIDIA DGX Spark (GB10, $3K) and AMD Ryzen AI MAX+ 395 / X2 ($2K). Our findings demonstrate that multi-agent AI orchestration achieving 200-334 tok/s aggregate throughput is viable on accessible hardware, validating the "AGI for the home" thesis. We introduce the Harmonic Stack Launcher, an auto-configuring deployment tool that optimizes parallel slot allocation based on hardware detection.

1. Introduction

Traditional AI deployment assumes cloud infrastructure or expensive enterprise hardware. Our research explores whether consumer-grade unified memory architectures can support multi-agent AI systems at scale.

The Harmonic Stack is a tiered multi-agent architecture where specialized models collaborate on complex tasks. Efficient deployment requires maximizing aggregate throughput while respecting memory constraints--a optimization problem that varies significantly across hardware platforms.

1.1 Research Questions

1. How does parallel inference scale on unified memory architectures?
2. What are the optimal parallelism settings for each platform?
3. Can consumer hardware achieve sufficient throughput for real-time multi-agent orchestration?

2. Hardware Platforms

2.1 NVIDIA DGX Spark (SPARKY)

Specification	Value
SoC	NVIDIA GB10
Memory	128GB unified LPDDR5X
Memory Bandwidth	~500 GB/s
TDP	45-77W observed
Price	~$3,000

2.2 AMD Ryzen AI MAX+ 395 (ARCY)

Specification	Value
CPU	Zen 5, 16 cores
GPU	Radeon 8060S (RDNA 3.5)
Memory	128GB unified DDR5
GPU Allocation	92GB (BIOS configured)
Memory Bandwidth	~256 GB/s
Price	~$2,000

3. Methodology

3.1 Test Configuration

Both systems ran Ollama 0.15.x with identical settings:

```
OLLAMA_NUM_PARALLEL=64
```

For AMD:

```
HSA_OVERRIDE_GFX_VERSION=11.0.0
```

3.2 Benchmark Protocol

- Models: qwen3:4b (2.5GB), qwen3:8b (5.2GB), Harmonic Stack agents (5GB each)
- Parallelism levels: 1x, 2x, 4x, 8x, 12x, 16x, 20x, 24x, 32x
- Trials: 3 per configuration
- Prompt: 50 tokens input, ~100 tokens output
- Metric: Aggregate tokens/second across all parallel streams

3.3 Critical Discovery: NUM_PARALLEL Setting

Initial X2 results showed flat scaling (51->61 tok/s from 1x->8x). Investigation revealed OLLAMA_NUM_PARALLEL defaults to 1, serializing all requests regardless of hardware capability.

Setting OLLAMA_NUM_PARALLEL=64 transformed X2 scaling from 1.2x to 5x throughput gain at 8x parallelism.

4. Results

4.1 Small Model Comparison (qwen3:4b, 2.5GB)

Parallelism	ARCY (X2)	SPARKY (DGX)	Ratio
1x	27.7 tok/s	21.7 tok/s	1.28x
2x	39.1 tok/s	35.9 tok/s	1.09x
4x	54.2 tok/s	35.2 tok/s	1.54x
8x	69.3 tok/s	119.4 tok/s	0.58x
12x	**222.6 tok/s**	151.1 tok/s	1.47x
16x	163.4 tok/s	222.9 tok/s	0.73x
20x	-	258.6 tok/s	-
24x	-	266.5 tok/s	-
32x	-	**308.2 tok/s**	-

4.2 DGX Spark Full Results

Small Models (~2.5GB)

Model	Peak Parallel	Peak tok/s	Efficiency
qwen3:4b	32x	308.2	7.87 tok/W
executive	16x	334.1	8.82 tok/W
operator	32x	285.3	6.73 tok/W

Medium Models (~5GB)

Model	Peak Parallel	Peak tok/s	Efficiency
qwen3:8b	16x	285.3	6.95 tok/W
technical_director	16x	248.2	5.66 tok/W
research_director	16x	252.6	6.20 tok/W
creative_director	16x	250.4	6.08 tok/W

Large Models (~9GB)

Model	Peak Parallel	Peak tok/s	Efficiency

qwen3:14b	16x	~120	~2.8 tok/W

4.3 Scaling Characteristics

![Scaling Curves](scaling_curves.png)

DGX Spark (GB10):

- Near-linear scaling to 16x
- Continues improving to 32x on small models
- Sweet spot: 16x parallel
- Peak: 334.1 tok/s (executive model)

X2 (Ryzen AI MAX+):

- Excellent scaling to 12x
- Performance cliff at 16x+
- Sweet spot: 12x parallel
- Peak: 222.6 tok/s (qwen3:4b)

4.4 Efficiency Analysis

Platform	Peak tok/s	Power	Efficiency
DGX Spark	334.1	39W	**8.82 tok/W**
X2	222.6	~65W*	~3.4 tok/W

*X2 power estimated; no direct measurement available

5. Harmonic Stack Launcher

Based on benchmark findings, we developed an auto-configuring deployment tool.

5.1 Hardware Detection

```
hardware = detect_hardware()
# Returns: {'profile': 'dgx_spark', 'gpu_mem_gb': 128, 'peak_parallel': 16}
```

5.2 Tier-Based Allocation

Models are allocated parallel slots by priority:

Tier	Role	Allocation

1	Executive	100% of peak_parallel
2	Directors	75% of peak_parallel
3	Specialists	50% of peak_parallel
4	Heavy	33% of peak_parallel

5.3 Memory Budget

```
Model RAM = base_weights + (num_parallel x kv_cache_per_slot)
```

Model	Base	KV/slot	@8x	@12x	@16x
qwen3:4b	2.5GB	0.3GB	4.9GB	6.1GB	7.3GB
qwen3:8b	5.2GB	0.5GB	9.2GB	11.2GB	13.2GB
qwen3:14b	9.3GB	0.8GB	15.7GB	18.9GB	22.1GB

5.4 Example Deployment

DGX Spark (128GB):

```
[Tier 1: EXECUTIVE]
  executive                  16x  (7.3GB)
  operator                   16x  (7.3GB)

[Tier 2: DIRECTORS]
  technical_director         12x  (11.2GB)
  research_director          12x  (11.2GB)
  creative_director          12x  (11.2GB)

[Tier 3: SPECIALISTS]
  coder                       8x  (15.7GB)
  analyst                     8x  (9.2GB)

Total: 73.1GB / 108.8GB available
```

6. Discussion

6.1 Architecture Implications

The distinct scaling curves suggest different optimal use cases:

DGX Spark: Optimized for batch processing and multi-agent orchestration. The ability to scale to 32x parallel makes it ideal for Harmonic Stack deployments where many agents operate simultaneously.

X2 (Ryzen AI MAX+): Optimized for interactive, low-latency workloads. Peaks earlier but achieves excellent single-stream performance. Ideal for personal AI assistants and real-time coding companions.

6.2 Memory Bandwidth Correlation

The 2:1 ratio in memory bandwidth (500 GB/s vs 256 GB/s) correlates with the parallel scaling ceiling difference (32x vs 12x). This suggests memory bandwidth is the primary bottleneck for parallel inference on unified memory architectures.

6.3 AGI for the Home

Both platforms achieve >200 tok/s aggregate throughput--sufficient for real-time multi-agent collaboration. At $2-3K price points, this validates the thesis that meaningful AI capability can be deployed on consumer hardware without cloud dependency.

7. Conclusion

Consumer unified memory platforms are viable for multi-agent AI deployment:

1. DGX Spark achieves 334 tok/s peak at 16x parallel, 8.82 tok/W efficiency
2. X2 achieves 223 tok/s peak at 12x parallel, ~3.4 tok/W efficiency
3. Critical setting: OLLAMA_NUM_PARALLEL=64 unlocks true parallel scaling
4. Tier-based allocation optimizes memory budget across model priorities

The Harmonic Stack Launcher automates deployment configuration, enabling accessible AI orchestration on home hardware.

8. Availability

- Harmonic Stack Launcher: https://github.com/joehoeller/harmonic-stack
- Benchmark Data: https://github.com/joehoeller/harmonic-stack/tree/main/benchmarks
- Pre-built Models: ollama pull ghostinthemachine/harmonic-stack

References

1. Ollama Documentation. https://ollama.ai/docs
2. NVIDIA DGX Spark Technical Specifications
3. AMD Ryzen AI MAX+ Product Brief

Ghost in the Machine Labs

501(c)(3) operating under "All Watched Over By Machines Of Loving Grace"