Scalability Analysis Report

*Report generated: 2025-07-09 08:57:05*

# Executive Summary

This report analyzes the scalability characteristics of the system under test across 4 different resource levels (from 2 to 16).

## Key Findings

* Maximum Throughput: 353.28 requests/sec achieved at resource level 8
* Best Response Time: 56.05 ms achieved at resource level 8
* Maximum Speedup: 1.98x achieved at resource level 8 compared to baseline

# Detailed Performance Metrics

|  |  |  |  |
| --- | --- | --- | --- |
| **Resource Level** | **Throughput (req/s)** | **Avg Response Time (ms)** | **Error %** |
| 2 | 178.83 | 111.59 | 1.00 |
| 4 | 275.63 | 70.34 | 1.00 |
| 8 | 353.28 | 56.05 | 1.00 |
| 16 | 330.93 | 60.92 | 1.00 |

# Basic Scalability Metrics

|  |  |  |
| --- | --- | --- |
| **Resource Level** | **Throughput Speedup** | **Scalability Efficiency** |
| 2 | 1.00x | 100.00% |
| 4 | 1.54x | 77.07% |
| 8 | 1.98x | 49.39% |
| 16 | 1.85x | 23.13% |

# Advanced Scalability Analysis

## Amdahl's Law Analysis

**Parallelizable portion:** 50.14%

**Serial portion:** 49.86%

**Theoretical maximum speedup:** 2.01x

## Gustafson's Law Analysis

**Scalable portion:** 7.47%

**Fixed portion:** 92.53%

## Universal Scalability Law Analysis

**Contention factor (σ):** 0.4986

**Coherency factor (κ):** 0.0000

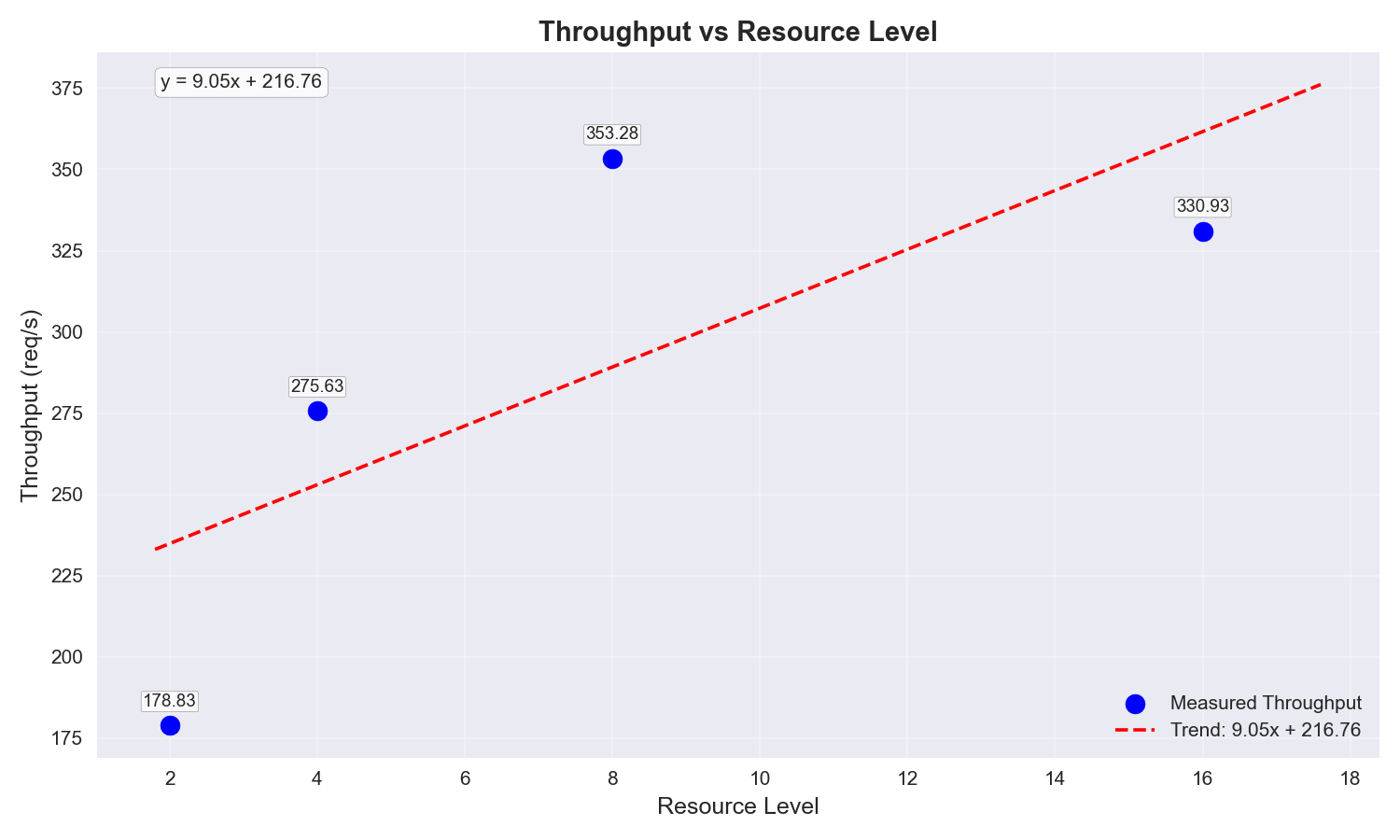
## Model Interpretations

* **Amdahl:** Moderate parallelizability. The system can benefit from additional resources, but with diminishing returns.
* **Gustafson:** Poor scalability with problem size. The system has significant fixed overhead that limits scaling with workload growth.
* **Usl:** Moderate scalability. Noticeable contention and some coherency delays. Scaling will be limited.

# Visual Analysis

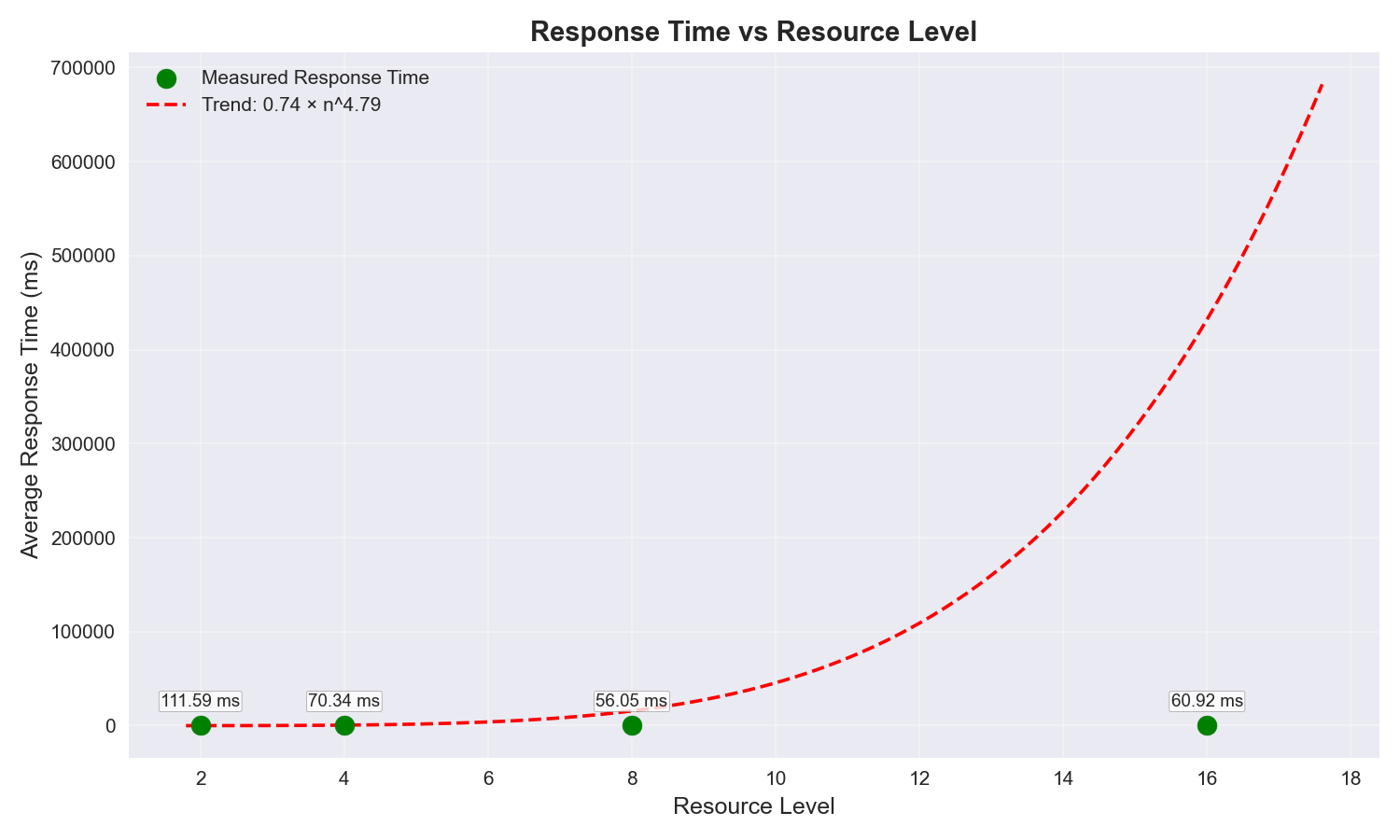
The following plots provide visual representation of the scalability characteristics.

## Throughput vs. Resource Level



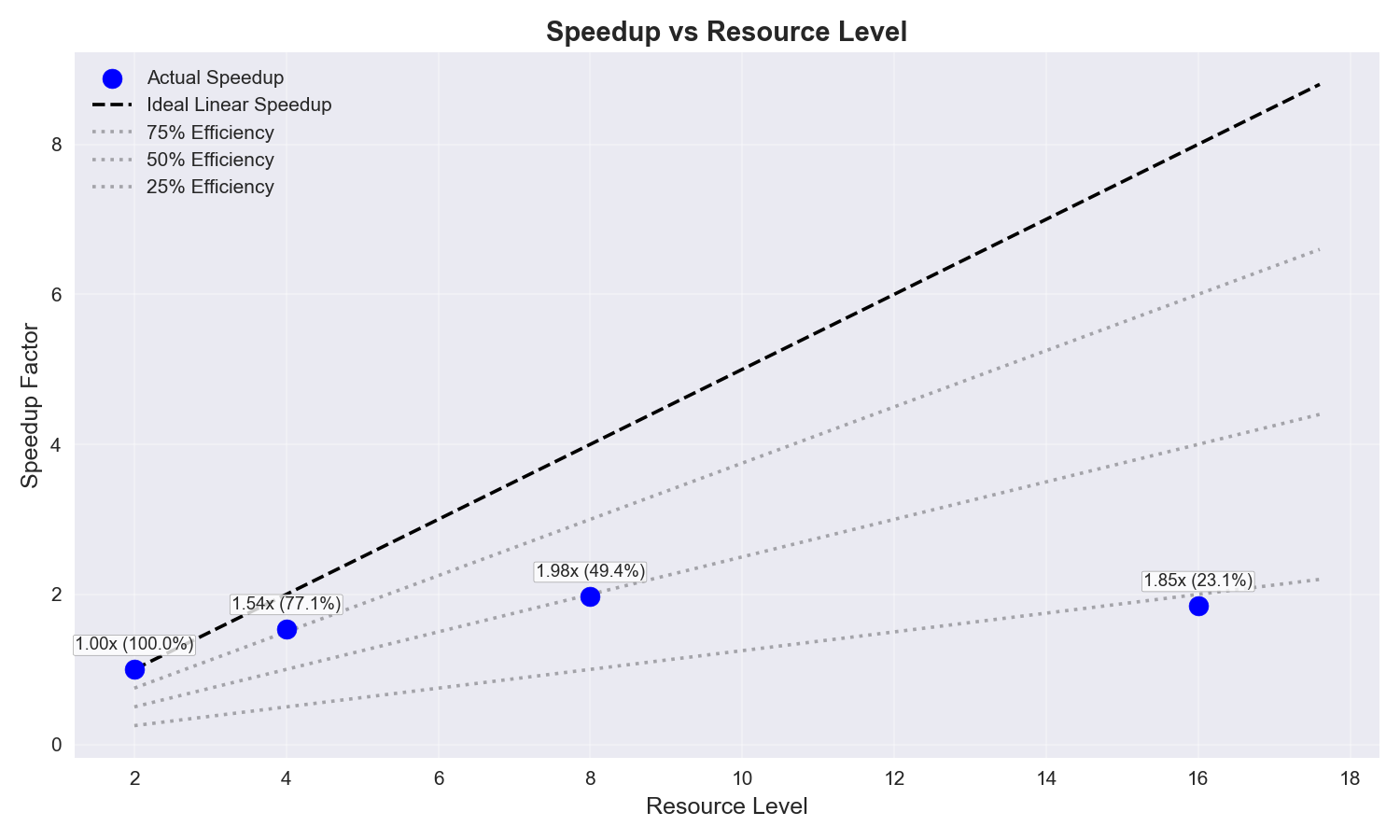
This plot shows how the system throughput changes as resources are added. The trend line indicates the overall scaling pattern.

## Response Time vs. Resource Level



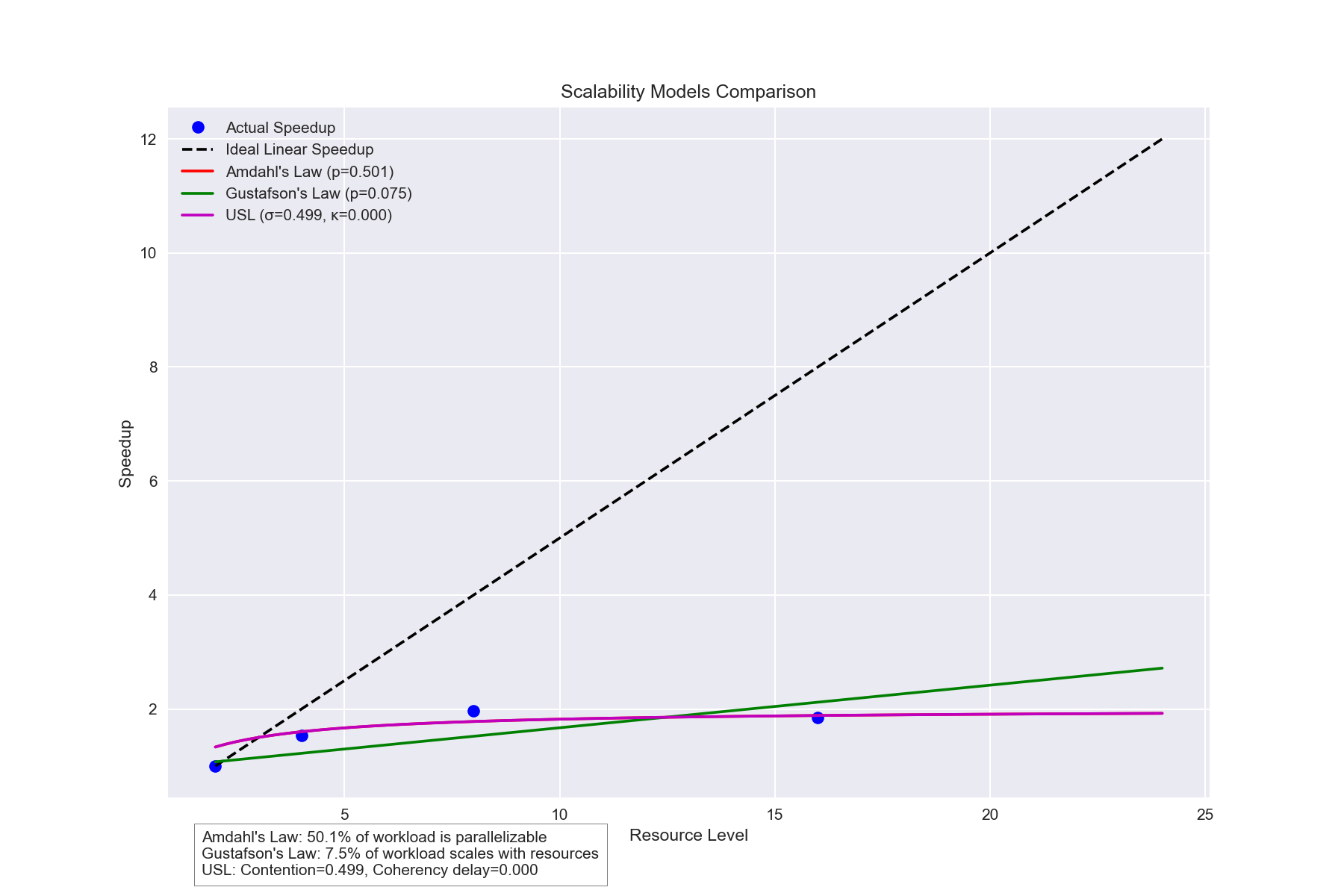
This plot illustrates how response times are affected by resource scaling. Lower values indicate better performance.

## Speedup vs. Resource Level



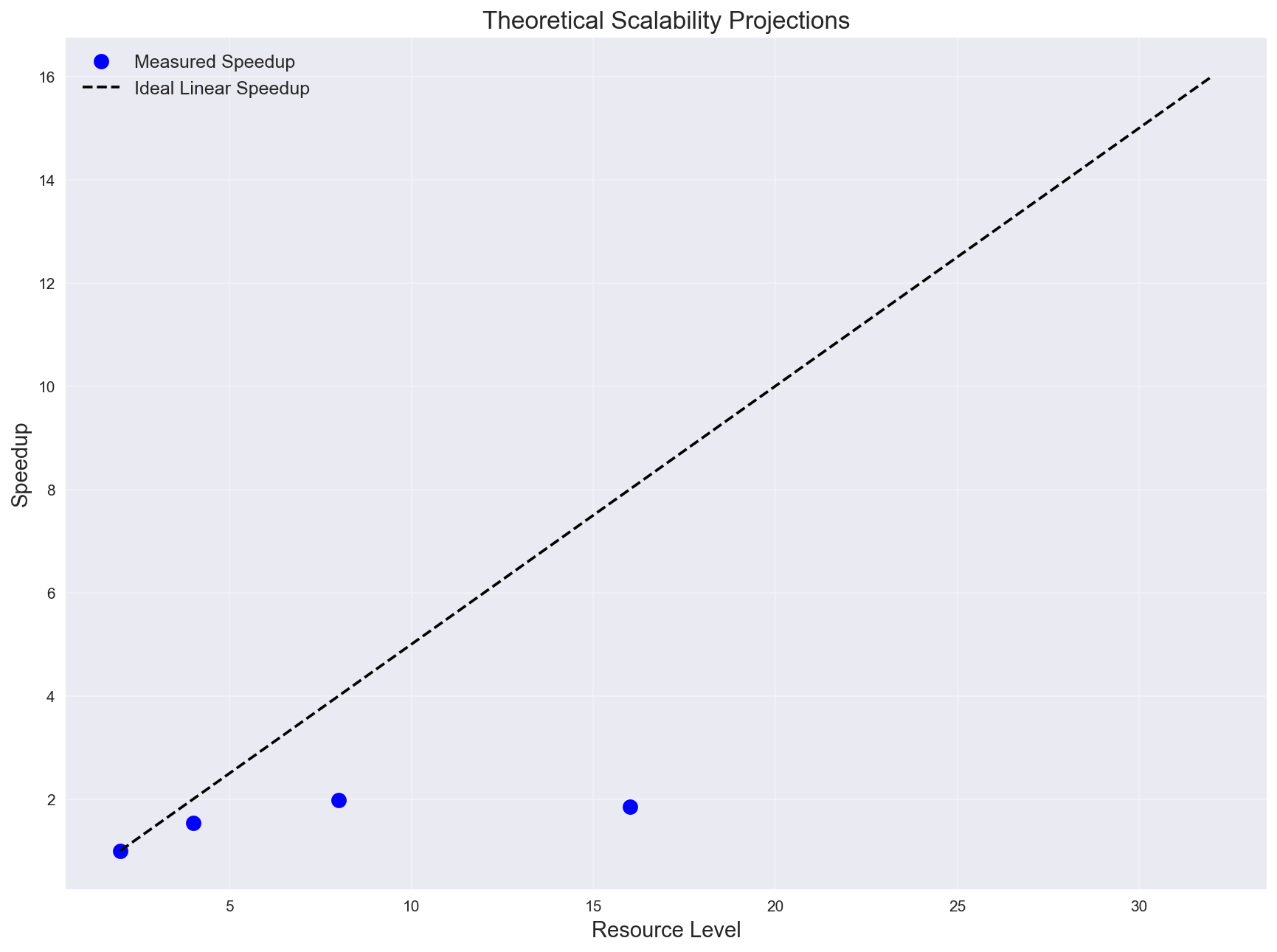
This plot compares actual speedup against ideal linear speedup. The gap between actual and ideal lines indicates efficiency loss as resources scale.

## Scalability Models Comparison



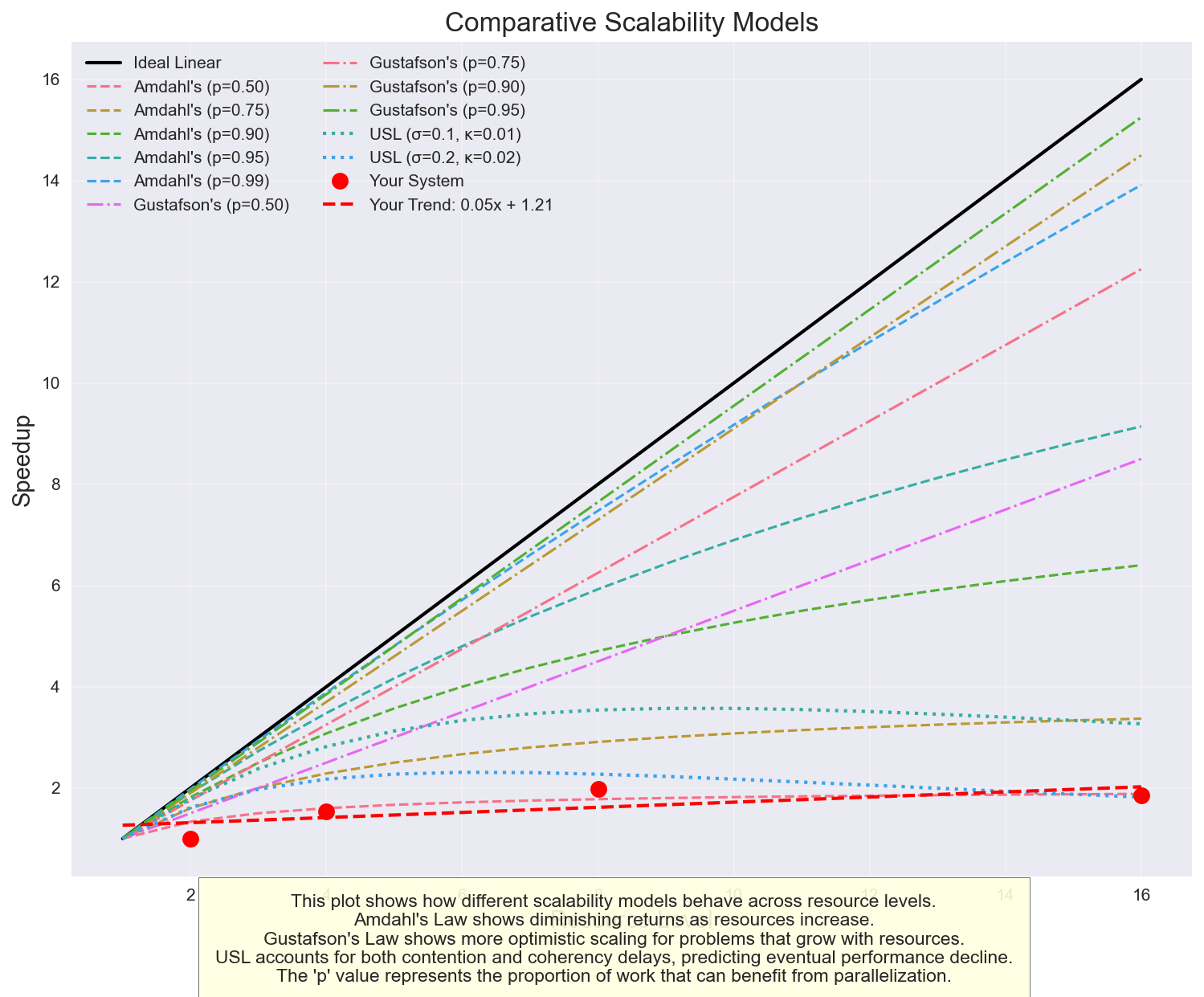
This plot compares the actual speedup with predictions from different scalability laws. The closest model to actual data points indicates which theoretical model best describes the system's scaling behavior.

## Theoretical Scalability Projections



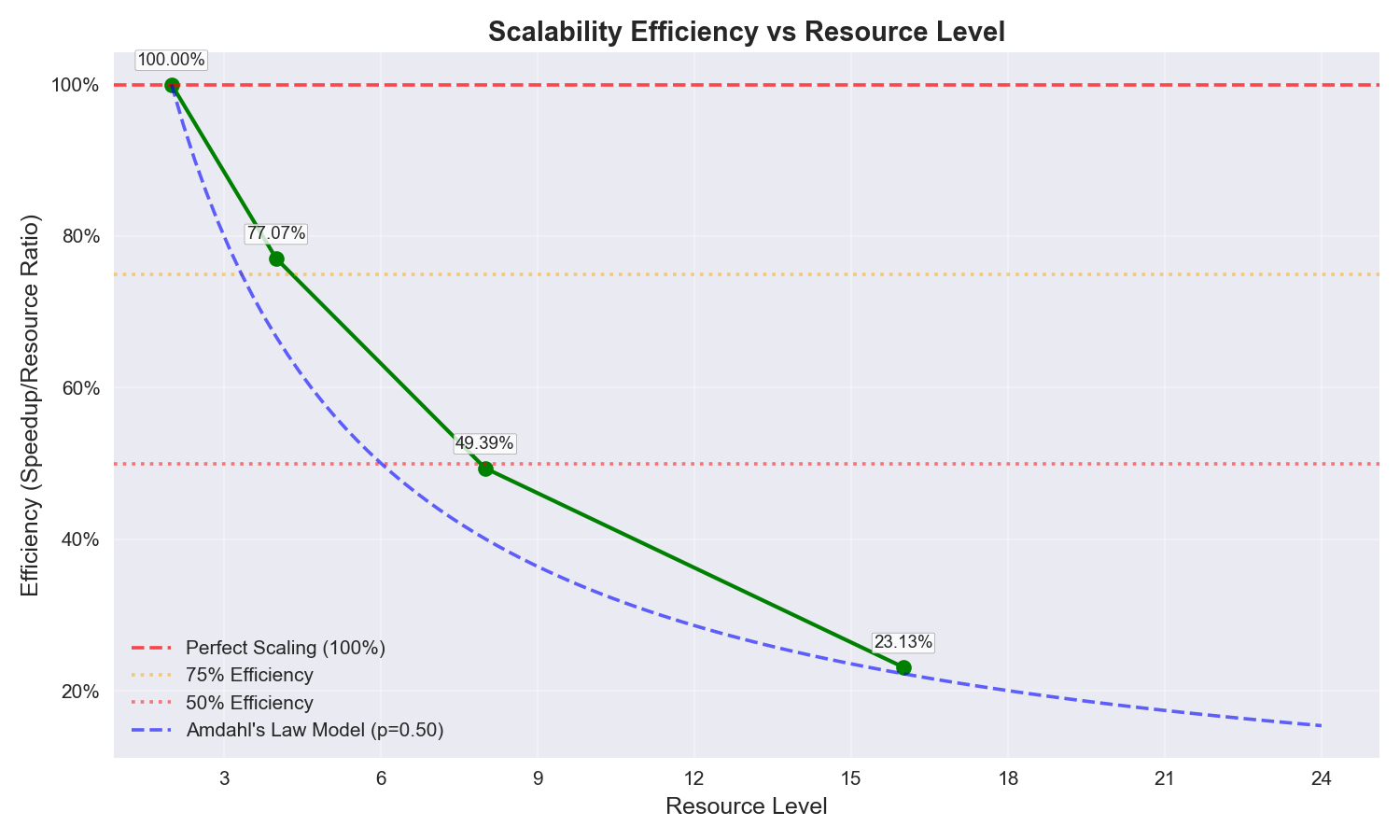
This plot shows theoretical projections of different scalability models based on observed data. It predicts how the system might scale with additional resources beyond those tested.

## Comparative Scalability Model Characteristics



This educational plot illustrates the fundamental differences between various scalability models with different parameters. It helps identify which theoretical model best describes your system's behavior.

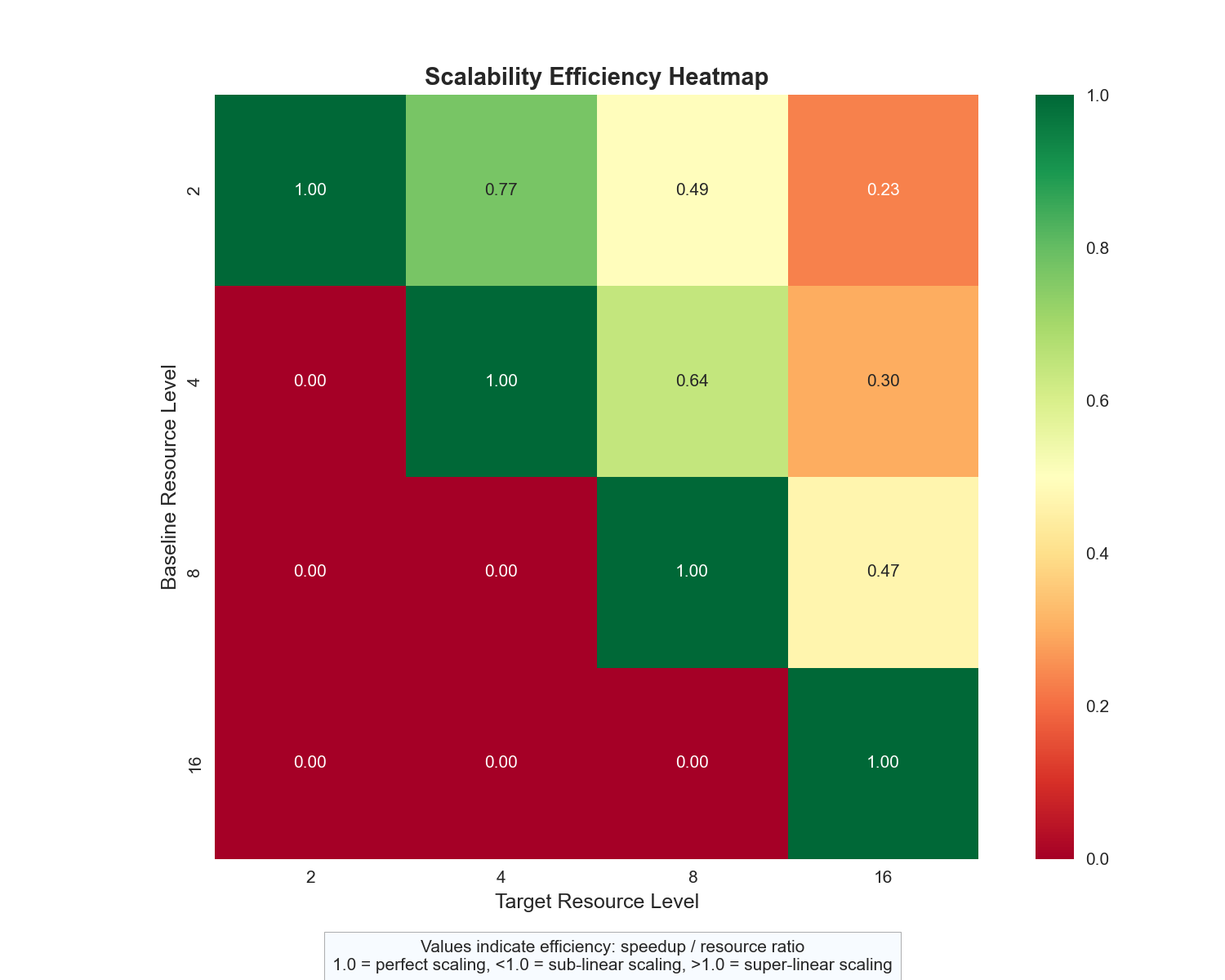
## Scalability Efficiency Analysis



This plot shows how efficiently your system scales as resources increase. The efficiency is calculated as speedup divided by resource ratio, with 100% representing perfect linear scaling. Declining efficiency indicates diminishing returns from additional resources.

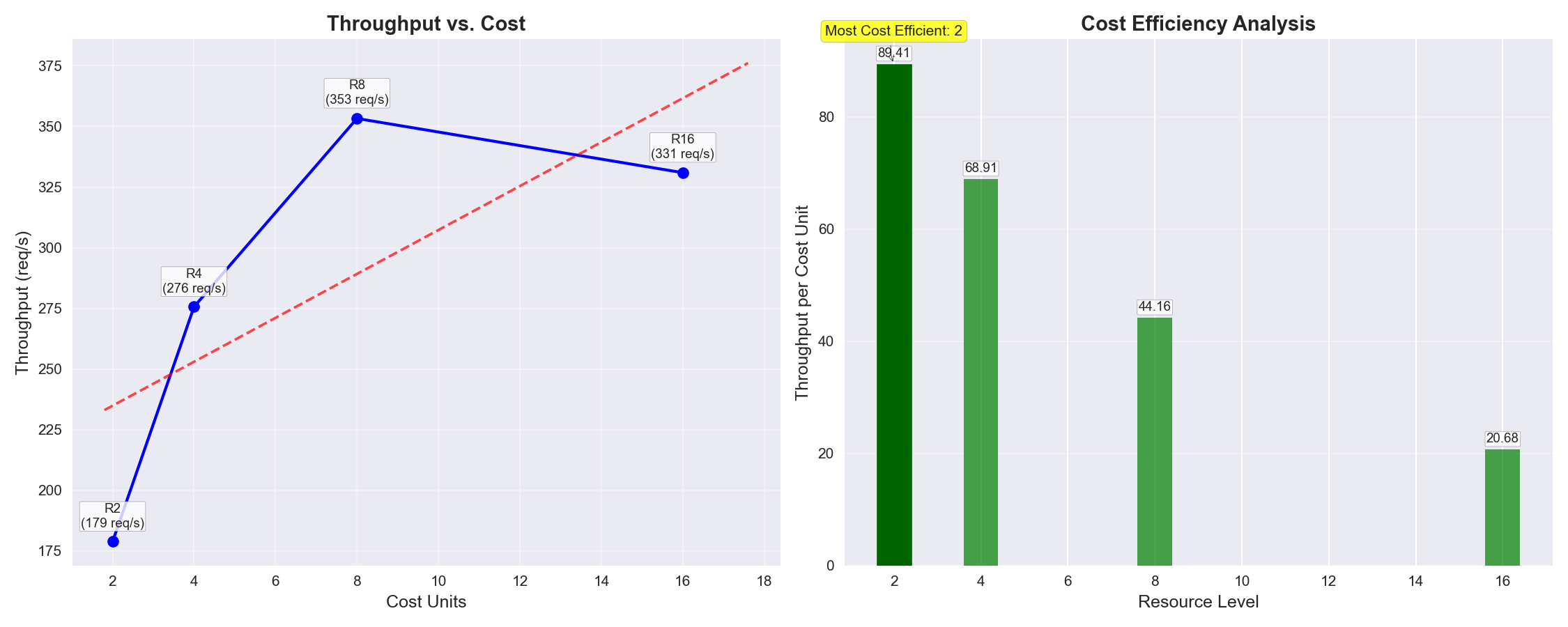
* **Current scaling efficiency:** 23.1% at 16 resources
* **Observation:** Lower scaling efficiency suggests significant serialization or contention in your system

## Efficiency Heatmap



This heatmap visualizes scaling efficiency between different resource levels. Each cell shows the efficiency when scaling from the baseline (row) to the target (column) resource level. Values close to 1.0 (green) indicate good scaling efficiency.

## Cost Efficiency Analysis



This dual visualization shows the relationship between cost and performance (left) and which configuration provides the best throughput per cost unit (right). The highlighted bar indicates the most cost-effective resource level for optimal return on investment.

* **Most cost-efficient configuration:** 2 resources
* **Recommendation:** The smallest configuration is most cost-efficient; consider if performance meets your needs

# Optimization Suggestions

* Consider optimizing the serial portions of the system, which account for 49.9% of execution time and limit maximum speedup to 2.0x.
* High contention factor (σ=0.499). Consider reducing shared resource access or implementing more efficient locking/synchronization mechanisms.

# Understanding Scalability Models

This section provides educational information about the different scalability models used in the analysis.

* **Amdahl's Law:** Shows diminishing returns as resources increase due to serial portions of work. Higher 'p' values indicate more parallelizable workloads and better scaling potential.
* **Gustafson's Law:** Shows more optimistic scaling when problem size grows with resources. Particularly relevant for workloads that can expand to use available resources.
* **Universal Scalability Law:** Accounts for both contention (σ) and coherency delays (κ). Predicts eventual performance decline at high resource levels due to increasing coordination costs.

**Based on your observed data points, your system currently exhibits:** 23.1% scaling efficiency, indicating poor scaling. This suggests major bottlenecks that severely limit parallel execution.