Scalability Analysis Report

*Report generated: 2025-07-15 15:06:10*

# Executive Summary

This report analyzes the scalability characteristics of the system under test across 5 different resource levels (from 1 to 6).

## Key Findings

* Maximum Throughput: 385.00 requests/sec achieved at resource level 6
* Best Response Time: 51.36 ms achieved at resource level 6
* Maximum Speedup: 3.69x achieved at resource level 6 compared to baseline

# Detailed Performance Metrics

|  |  |  |  |
| --- | --- | --- | --- |
| **Resource Level** | **Throughput (req/s)** | **Avg Response Time (ms)** | **Error %** |
| 1 | 104.36 | 200.07 | 1.00 |
| 2 | 185.11 | 106.81 | 1.00 |
| 3 | 257.02 | 76.98 | 1.00 |
| 4 | 306.90 | 63.07 | 1.00 |
| 6 | 385.00 | 51.36 | 1.00 |

# Basic Scalability Metrics

|  |  |  |
| --- | --- | --- |
| **Resource Level** | **Throughput Speedup** | **Scalability Efficiency** |
| 1 | 1.00x | 100.00% |
| 2 | 1.77x | 88.69% |
| 3 | 2.46x | 82.09% |
| 4 | 2.94x | 73.52% |
| 6 | 3.69x | 61.48% |

# Advanced Scalability Analysis

## Amdahl's Law Analysis

**Parallelizable portion:** 87.71%

**Serial portion:** 12.29%

**Theoretical maximum speedup:** 8.14x

## Gustafson's Law Analysis

**Scalable portion:** 58.89%

**Fixed portion:** 41.11%

## Universal Scalability Law Analysis

**Contention factor (σ):** 0.1050

**Coherency factor (κ):** 0.0034

**Optimal concurrency:** 132.12 resources

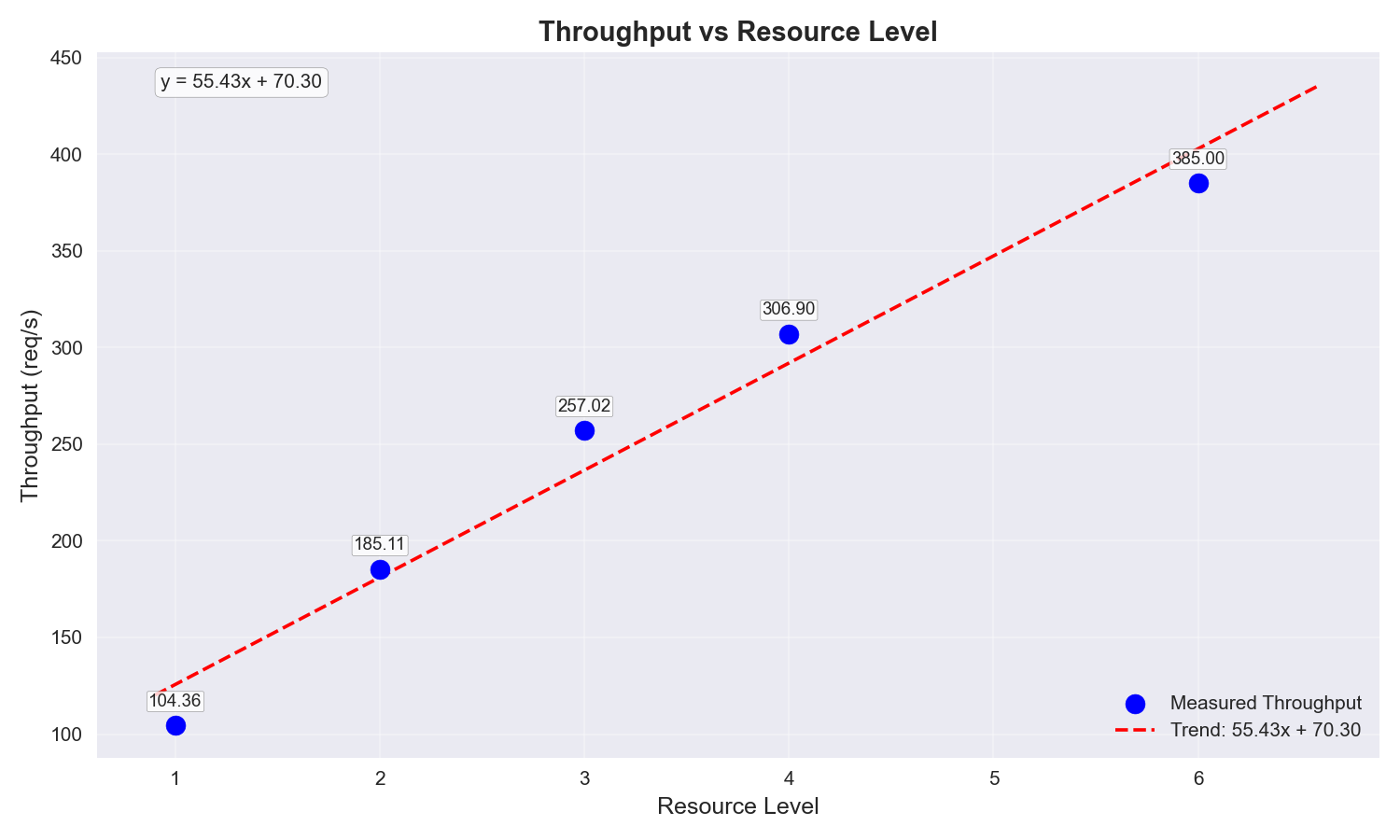
## Model Interpretations

* **Amdahl:** Good parallelizability. The system can benefit significantly from additional resources.
* **Gustafson:** Moderate scalability with problem size. The system has some fixed overhead that limits perfect scaling.
* **Usl:** Good scalability. Some contention but limited coherency issues. The system can scale reasonably well.

# Visual Analysis

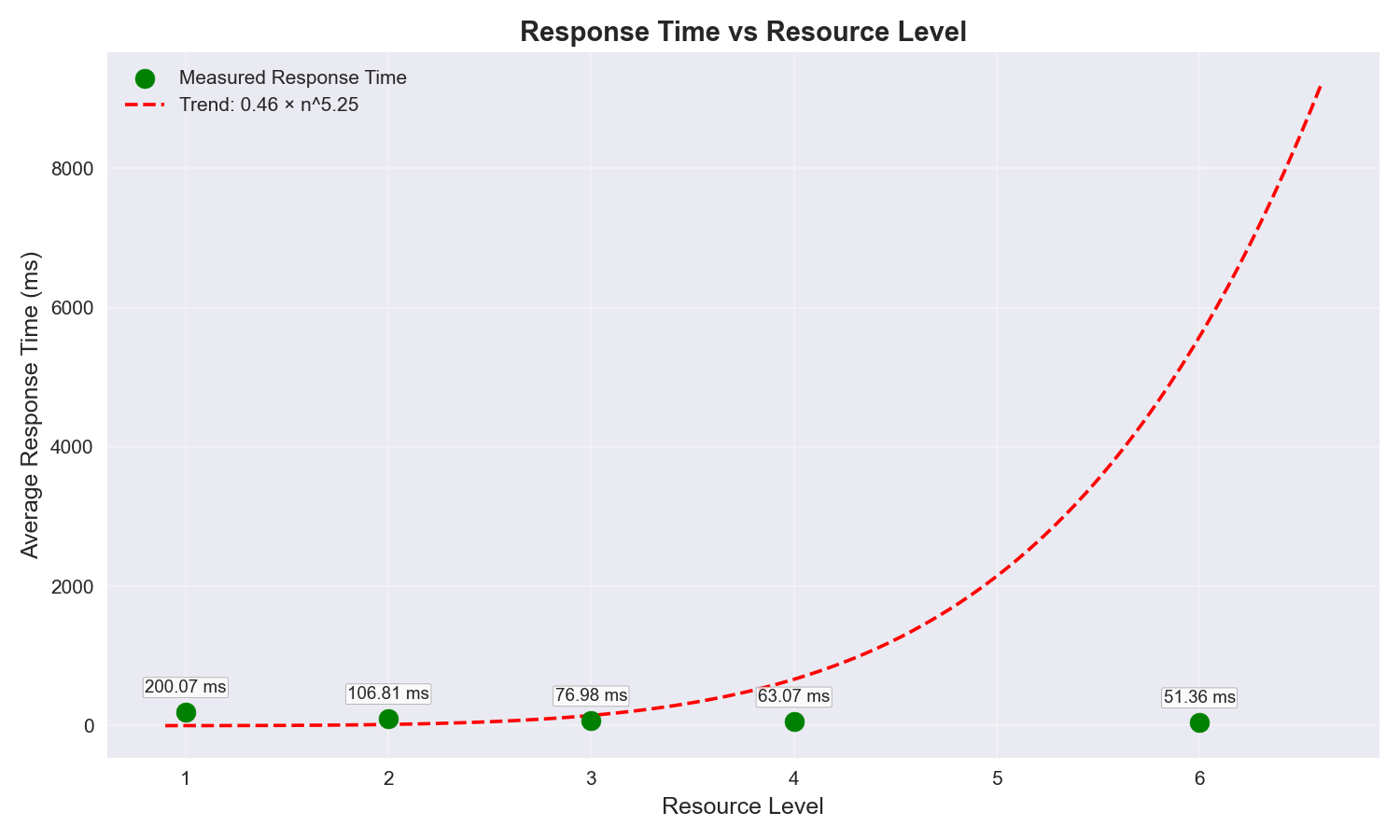
The following plots provide visual representation of the scalability characteristics.

## Throughput vs. Resource Level



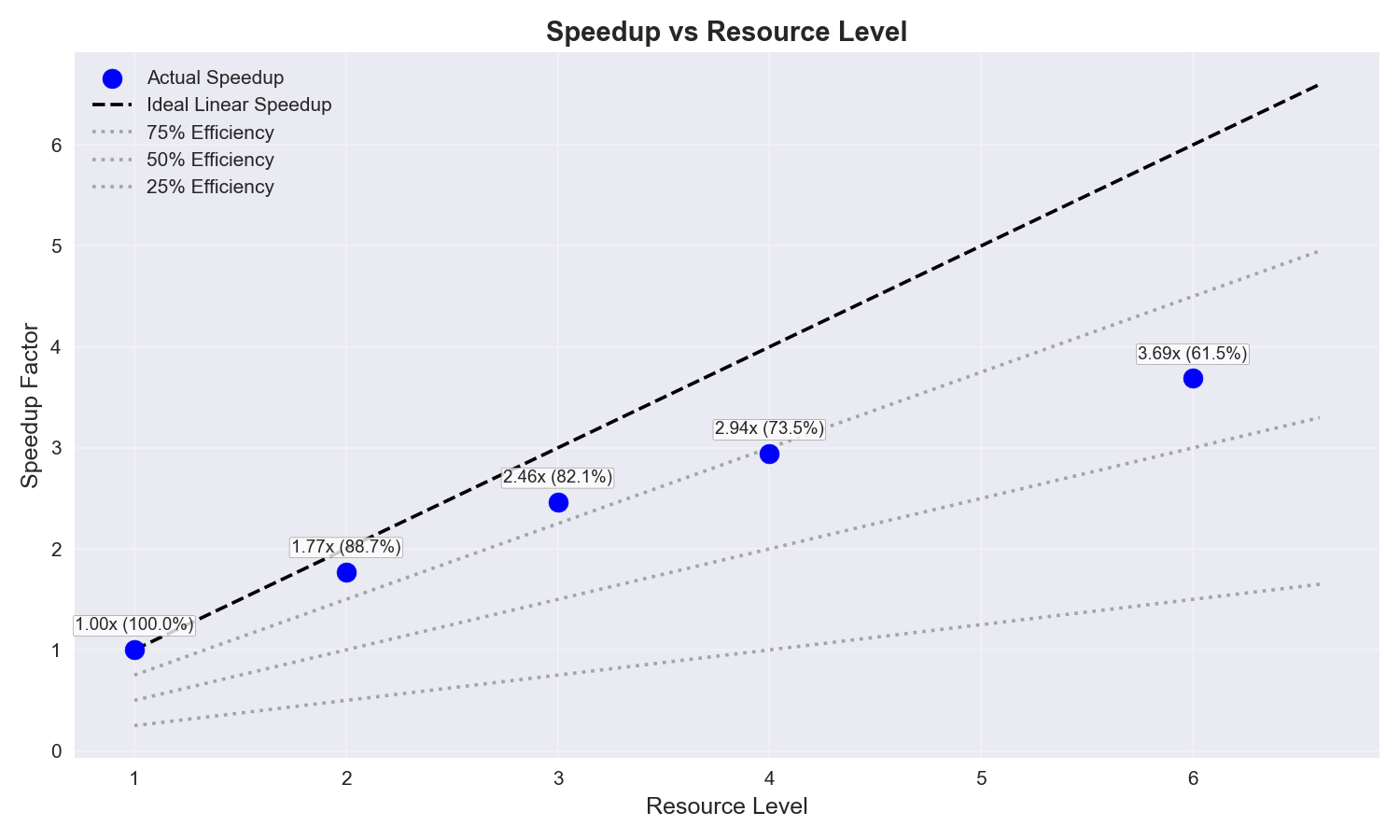
This plot shows how the system throughput changes as resources are added. The trend line indicates the overall scaling pattern.

## Response Time vs. Resource Level



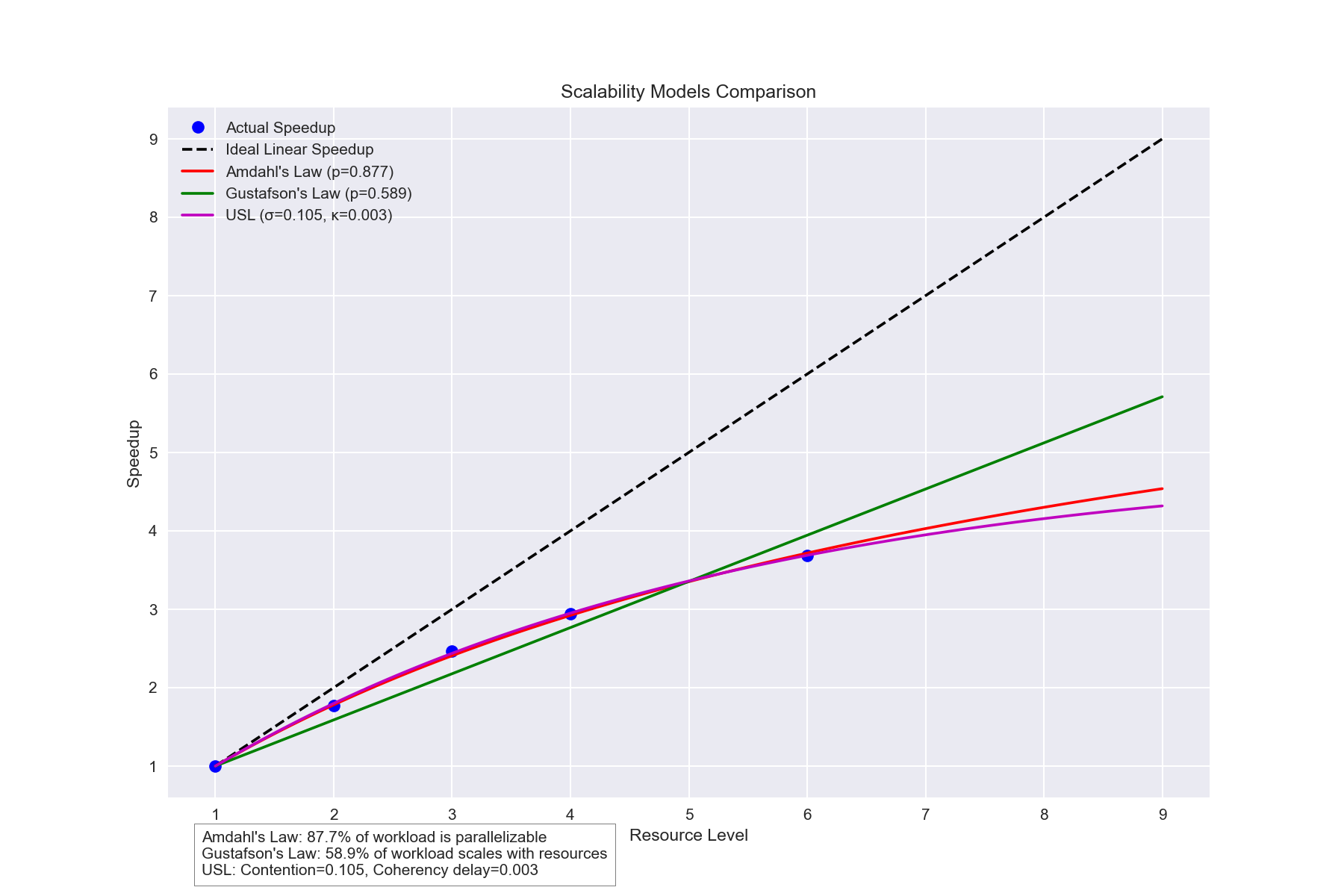
This plot illustrates how response times are affected by resource scaling. Lower values indicate better performance.

## Speedup vs. Resource Level



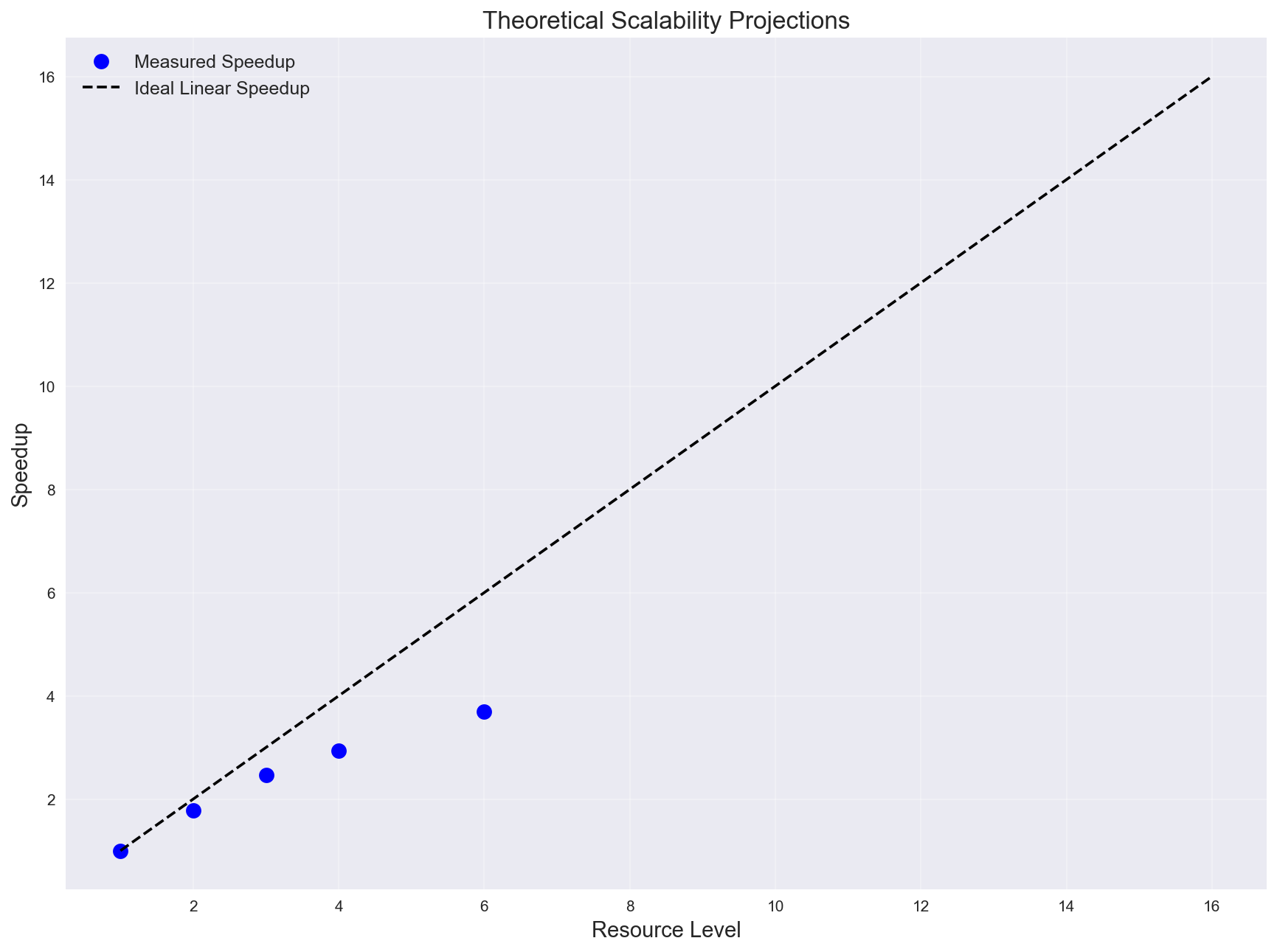
This plot compares actual speedup against ideal linear speedup. The gap between actual and ideal lines indicates efficiency loss as resources scale.

## 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



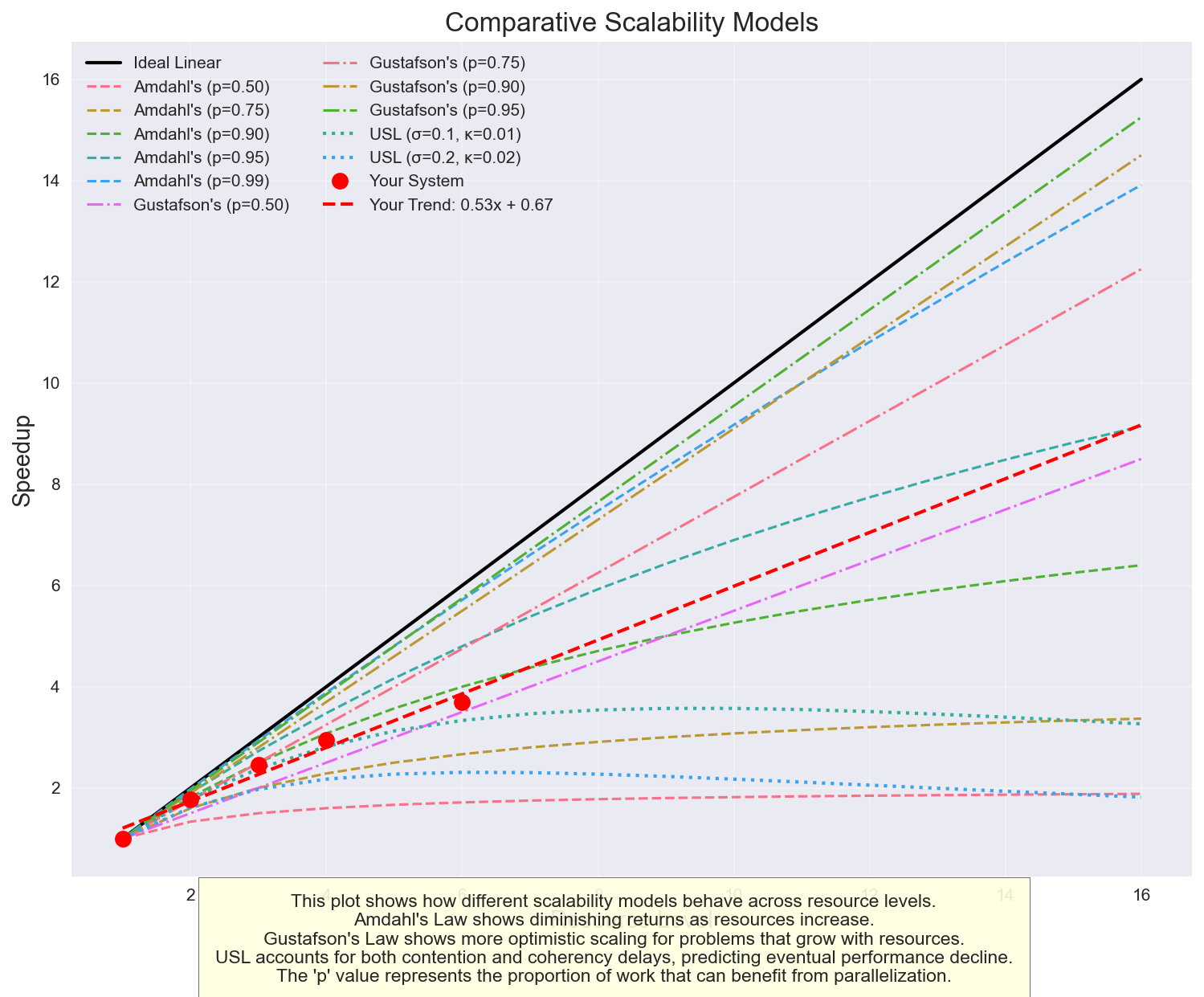
This plot shows theoretical projections of different scalability models based on observed data. It predicts how the system might scale with additional resources beyond those tested.

## 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 based on the shape of your actual data points.

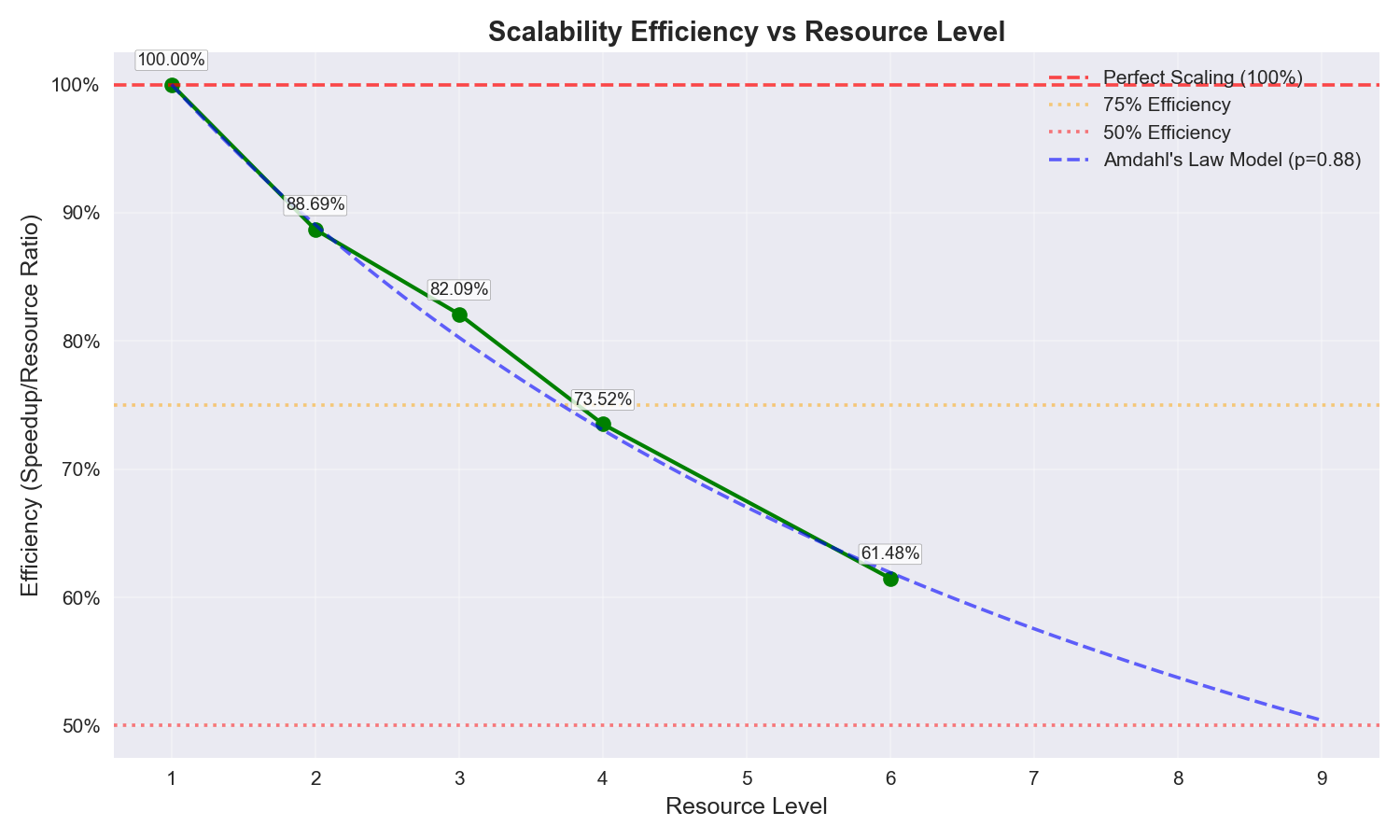
**Key model characteristics:**

* **Linear scaling:** Represents perfect scaling where adding resources results in proportional performance improvement.
* **Amdahl's Law:** Shows diminishing returns due to serial portions in the workload, with a clear asymptotic limit.
* **Gustafson's Law:** Exhibits better scaling than Amdahl's when workload size increases with resources.
* **Universal Scalability Law:** Shows initial scaling followed by performance degradation due to contention and coherency delays.



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.

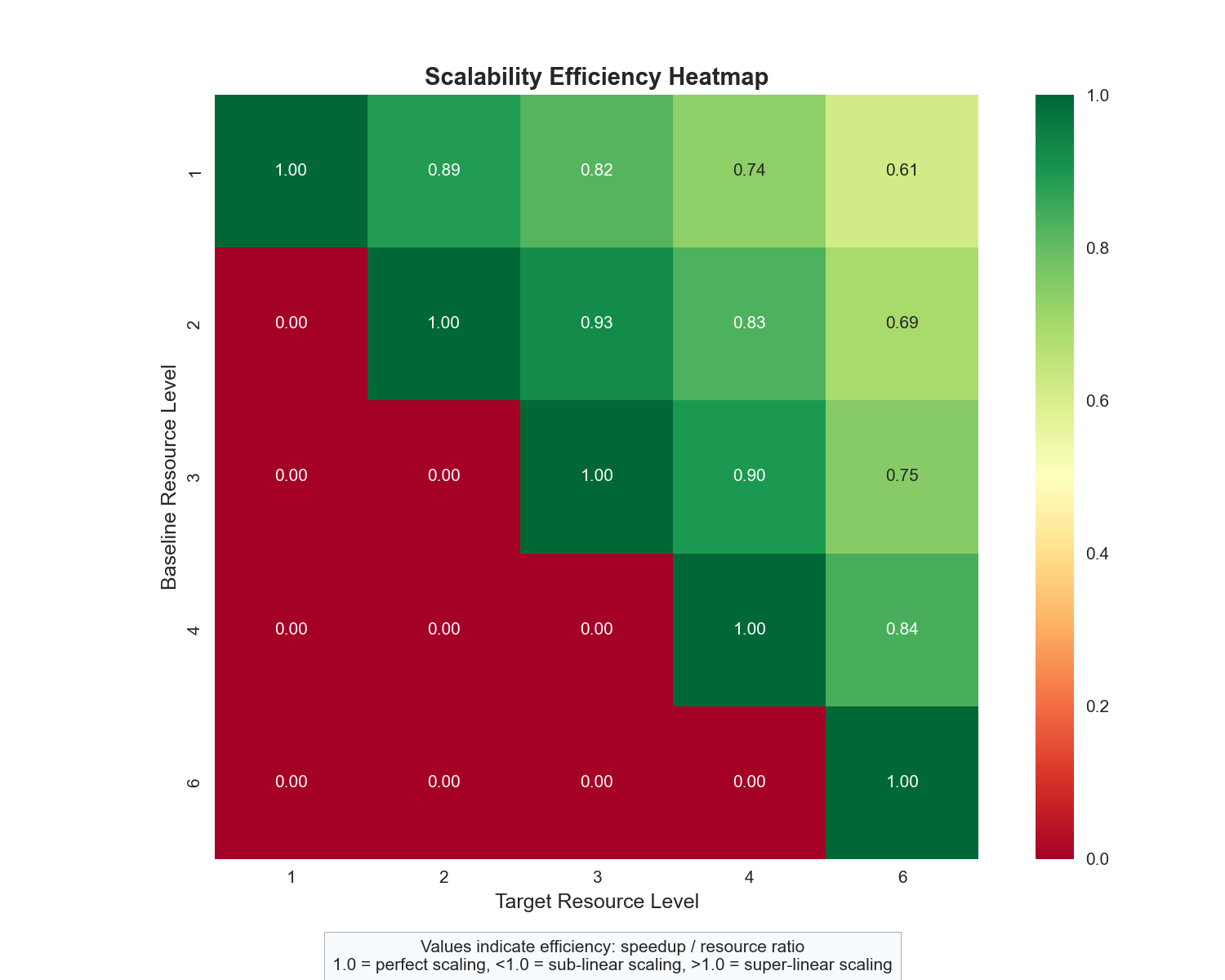
## Scalability Efficiency Analysis



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.

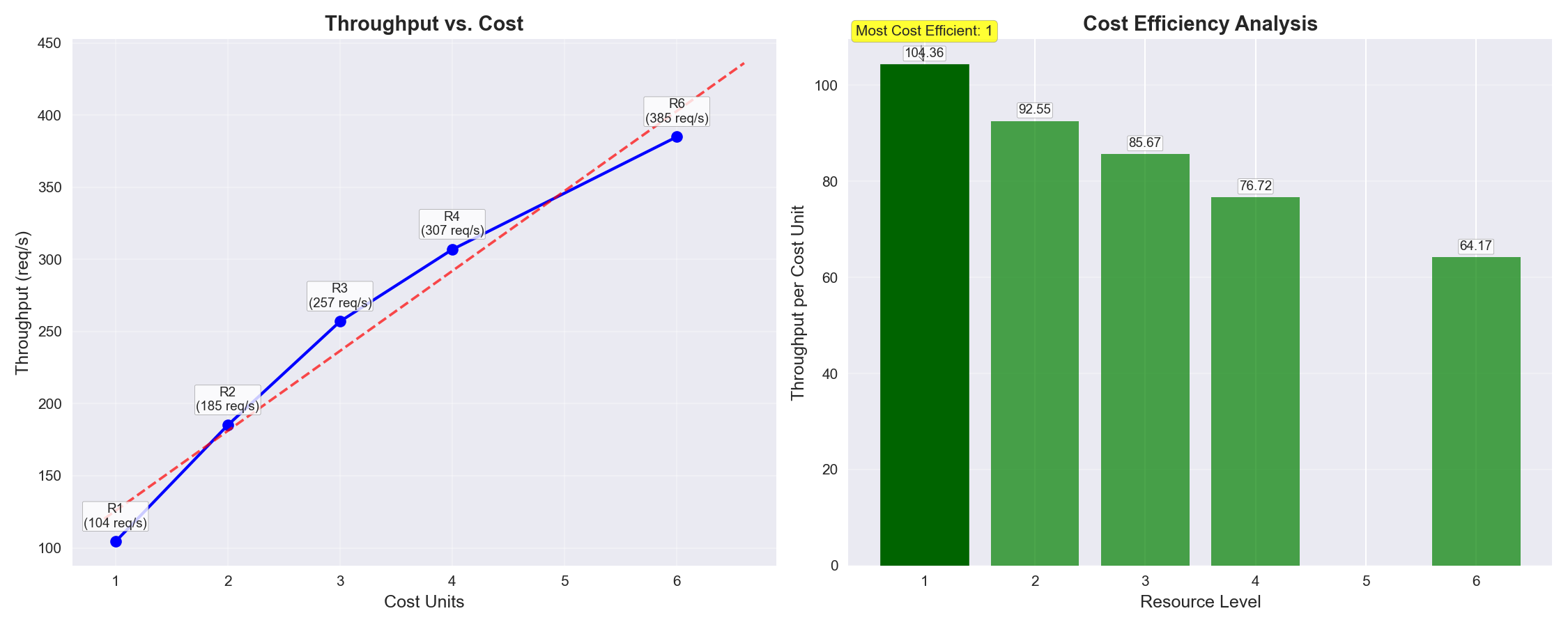
* **Current scaling efficiency:** 61.5% at 6 resources
* **Observation:** Lower scaling efficiency suggests significant serialization or contention in your system

## 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

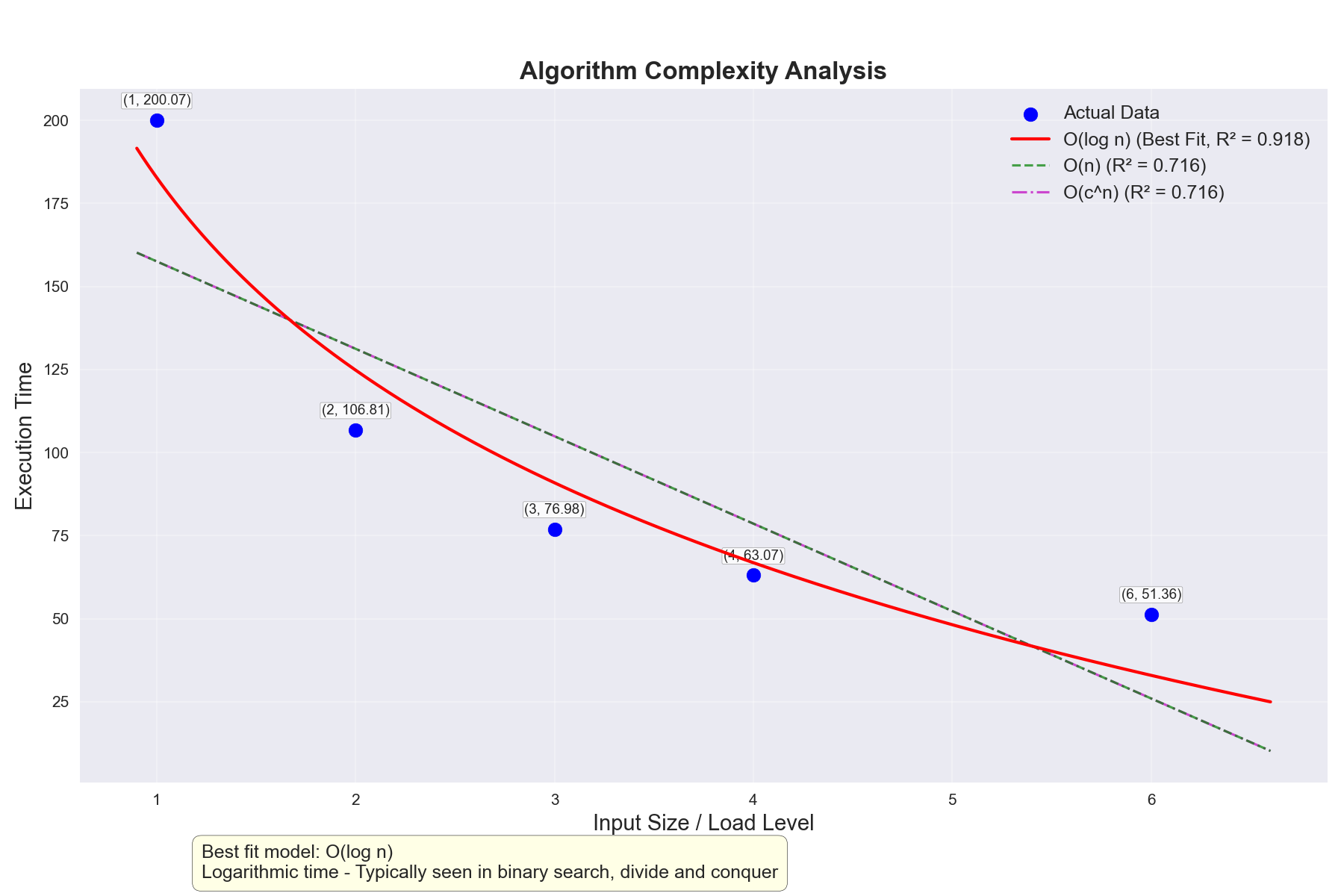


This dual visualization shows the relationship between cost and performance (left) and which configuration provides the best throughput per cost unit (right). The highlighted bar indicates the most cost-effective resource level for optimal return on investment.

**Why cost efficiency matters:** As you scale your system, it's crucial to understand not just raw performance gains but the cost-to-performance ratio. The linear cost model used in this analysis assumes costs scale directly with resource count.

* **Most cost-efficient configuration:** 1 resources with 104.36 throughput per cost unit
* **Cost efficiency comparison:** The most cost-efficient configuration is 162.6% more economical than using 6 resources
* **Recommendation:** For optimal resource utilization, consider using 1 resources when budget constraints are a priority.

## Algorithm Complexity Analysis



This plot shows the algorithmic complexity of your system by fitting various complexity models (O(1), O(log n), O(n), O(n log n), O(n²), etc.) to your performance data. The best fitting model indicates your system's computational complexity class, which influences how it will scale with increasing workloads.

**Understanding complexity classes:**

* **O(1) - Constant time:** Performance remains constant regardless of input size. Ideal for lookups and cached operations.
* **O(log n) - Logarithmic time:** Performance increases logarithmically with input size. Common in tree-based structures and binary searches.
* **O(n) - Linear time:** Performance scales linearly with input size. Typical in operations that process each input element once.
* **O(n log n) - Linearithmic time:** Performance scales slightly worse than linear. Common in efficient sorting algorithms.
* **O(n²) - Quadratic time:** Performance scales with the square of input size. Often indicates nested loops or inefficient algorithms.

**Why this matters for scalability:** The algorithmic complexity of your system directly impacts how it will handle larger workloads. Systems with better complexity classes (O(1), O(log n)) will generally scale better than those with higher complexity classes (O(n), O(n²)).

### Algorithm Complexity Interpretation

**Best fitting model:** O(log n)

**Confidence:** High confidence in the model fit.

**Explanation:** The system appears to have logarithmic time complexity.

**Implications:** This indicates very good scalability, typical of efficient search algorithms or data structures.

### Load Scalability Interpretation

**Saturation point:** 6 users/requests

**Optimal load point:** 6 users/requests

**Performance degradation observed:** No

#### Recommendations

* T
* h
* i
* s
* i
* s
* a
* l
* r
* e
* a
* d
* y
* v
* e
* r
* y
* e
* f
* f
* i
* c
* i
* e
* n
* t
* ;
* f
* o
* c
* u
* s
* o
* n
* o
* t
* h
* e
* r
* a
* s
* p
* e
* c
* t
* s
* o
* f
* o
* p
* t
* i
* m
* i
* z
* a
* t
* i
* o
* n
* i
* f
* n
* e
* e
* d
* e
* d
* .

# Optimization Suggestions

* The system performance is predicted to peak at 132.1 resources and decline with additional resources. Consider limiting deployment to this size.

# 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:** 61.5% scaling efficiency, indicating moderate scaling with significant overhead. This suggests substantial serial portions or contention limiting scalability.

## Key Findings

### Performance Metrics

Maximum Throughput: 385.00 requests/sec at 6 nodes

Best Response Time: 51.36 ms at 6 nodes

Maximum Speedup: 3.69x at 6 nodes (compared to baseline)

Scalability Efficiency: 61.48% at 6 nodes

### Scalability Model Parameters

Amdahl's Law: 87.71% parallelizable portion, 12.29% serial portion

Gustafson's Law: 58.89% scalable portion, 41.11% fixed portion

Universal Scalability Law:

- Contention factor (σ): 0.1050

- Coherency factor (κ): 0.0034

- Optimal concurrency: 132.12 nodes

### Algorithm Complexity

Best fitting model: O(log n)

Confidence: High

### Load Scalability

Saturation point: 6 users/requests

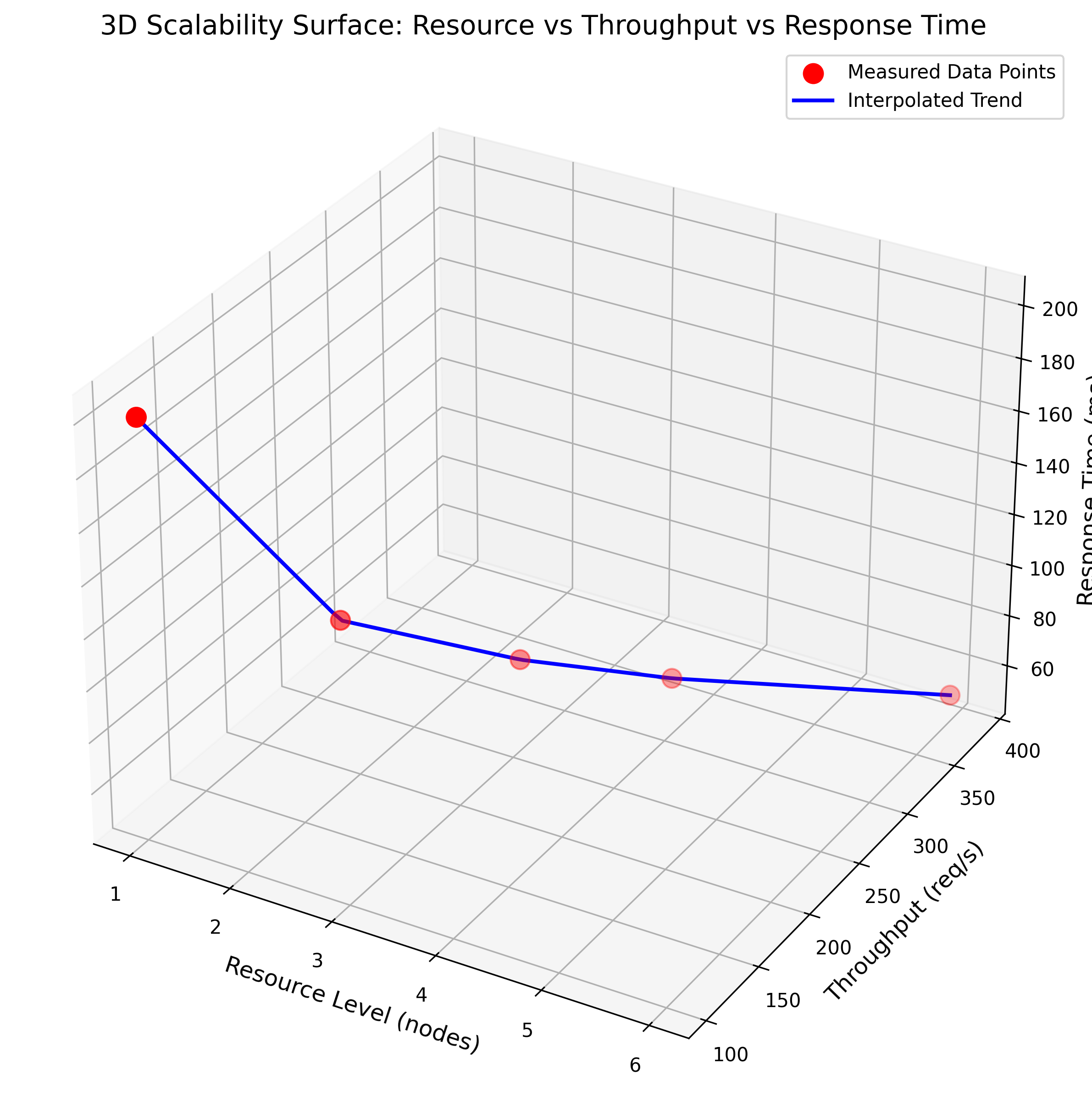
Optimal load point: 6 users/requests

Performance degradation: None observed

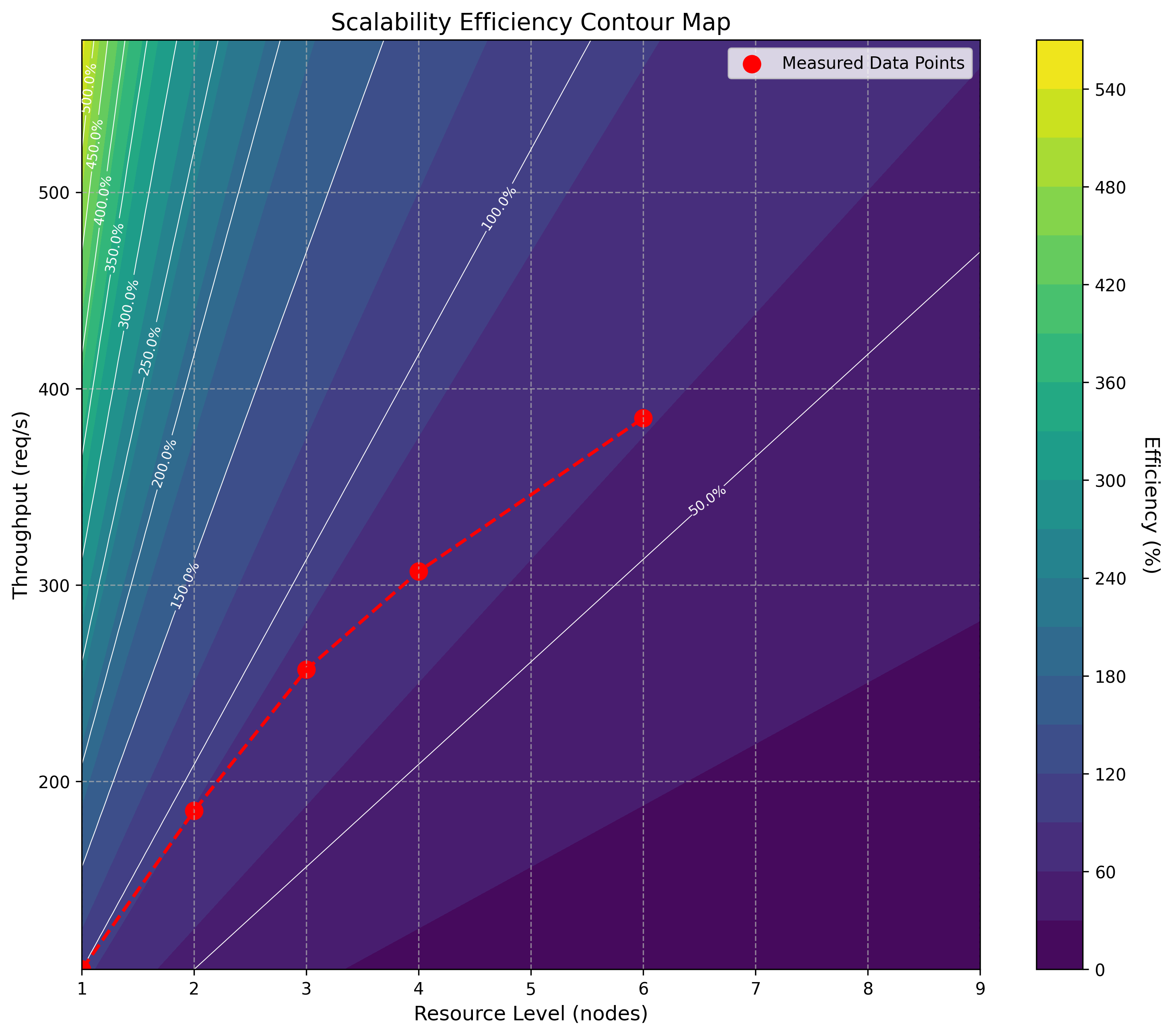
## Enhanced Visualizations

The following advanced visualizations provide additional insights into the system's scalability characteristics:

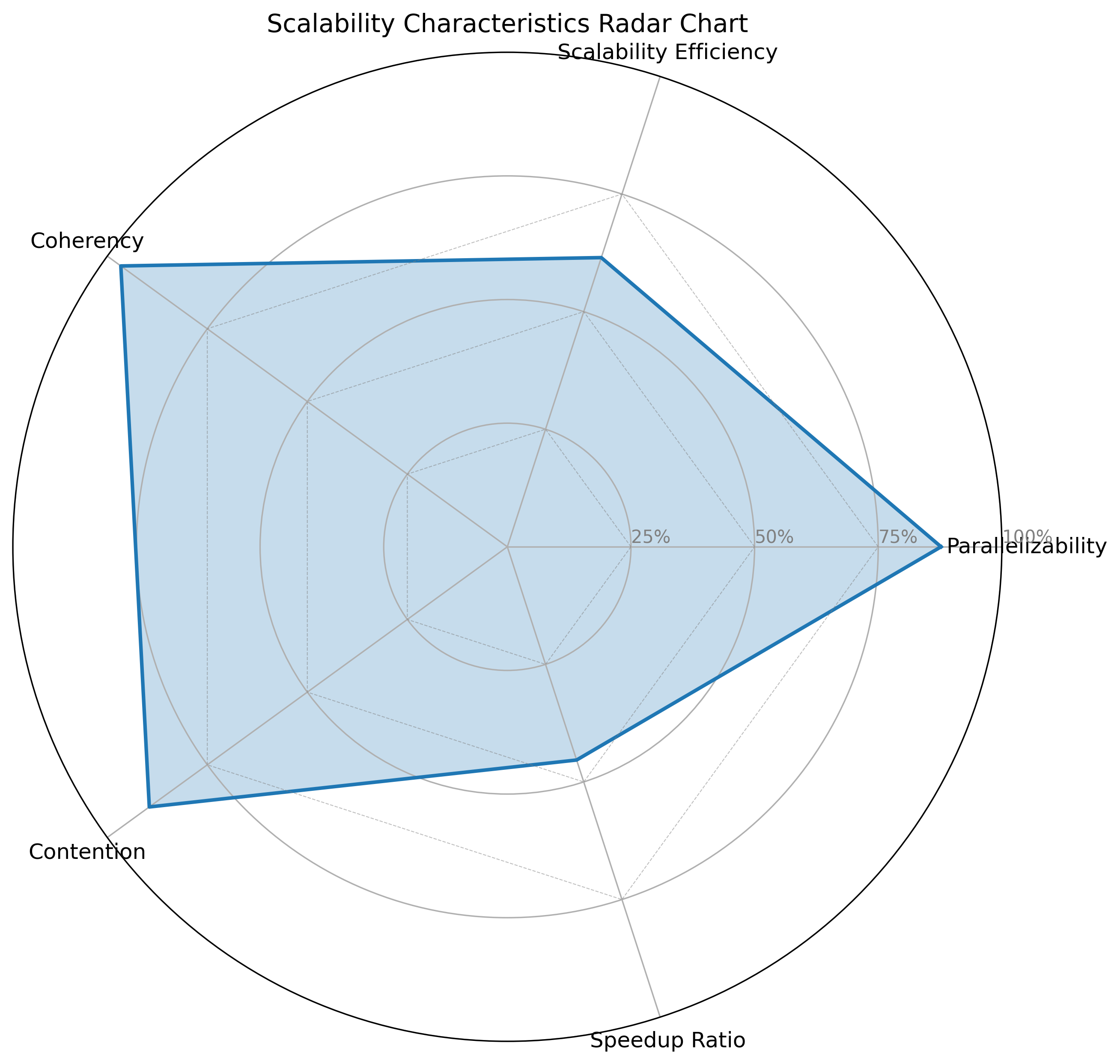
### 3D Scalability Surface



### Efficiency Contour Map



### Scalability Radar Chart



### Resource Optimization Chart

