Scalability Analysis of Bayesian Inference for Neutron Star Equation of State

Executive Summary

This report presents a comprehensive scalability analysis of our Bayesian inference workflow for neutron star equation of state (EOS) investigations using the UltraNest sampler with MPI parallelization. The study was conducted on the Deucalion HPC platform in Portugal, utilizing up to 22 compute nodes and 2,816 CPU cores. Our results demonstrate strong but sub-linear scaling, with performance improvements diminishing as the node count increases.

Introduction

Modern astrophysical investigations of neutron star properties require sophisticated Bayesian inference frameworks to constrain the equation of state from observational data. These computationally intensive workflows benefit significantly from high-performance computing resources but understanding their scaling behavior is crucial for efficient resource allocation.

Our implementation uses a Relativistic Mean Field (RMF) approach with UltraNest sampling to infer parameters of the neutron star equation of state, constrained by mass-radius observations. This report analyzes the performance scaling of this workflow as a function of computational resources.

Methodology

Computational Setup

- **HPC Platform**: Deucalion HPC platform (Portugal)
- **Nodes**: 2 to 22 compute nodes
- **CPU Cores**: 128 cores per node (256 to 2,816 cores total)
- **Inference Parameters**: 8 physical parameters (`g_sigma`, `g_omega`, `g_rho`, `kappa`, `lambda_0`, `zeta`, `Lambda_w`, `d1`)
- **Live Points**: 2,000
- **Maximum Function Calls**: 500,000
- **Test Code**: test_Inference_ultranest_mpi.py

The workflow implements a Bayesian inference using UltraNest with MPI parallelization, calculating the neutron star mass-radius relationship for various parameter combinations and comparing against observational constraints.

Results

Table 1 presents the performance metrics across different node configurations.

Table 1: Scalability Performance Metrics

Nodes	CPUs	Runtime (s)	Runtime (min)	Speedup	Efficiency
2	256	10,938.00	182.30	1.00	1.00
6	768	4,139.00	68.98	2.64	0.88
10	1,280	2,642.00	44.03	4.14	0.83
14	1,792	1,993.00	33.22	5.49	0.78
18	2,304	1,948.00	32.47	5.61	0.62
22	2,816	1,611.00	26.85	6.79	0.62

Where:

- **Speedup** = Runtime(2 nodes) / Runtime(N nodes)
- **Efficiency** = Speedup / (N nodes / 2 nodes)

Discussion

Performance Scaling

Our analysis reveals several key insights:

1. **Runtime Reduction**: Increasing the node count from 2 to 22 reduced the runtime from approximately 3 hours to under 30 minutes, representing a 6.79× speedup.

2. **Diminishing Returns**: The performance scaling shows clear diminishing returns. While adding nodes consistently improves absolute runtime, the marginal improvement decreases significantly past 14 nodes.

3. **Efficiency Drop**: Parallel efficiency decreases from 100% (by definition) at the baseline to 62% at 22 nodes, indicating growing communication overhead and load imbalance as the system scales.

4. **Performance Plateau**: Between 14 and 18 nodes, we observe a performance plateau with minimal runtime improvement (only 45 seconds or 2.3% reduction), suggesting a potential scaling bottleneck.

5. **Beyond 18 Nodes**: Performance improvement resumes beyond 18 nodes, though still with reduced efficiency.

Computational Bottlenecks

The observed scaling behavior suggests several potential bottlenecks:

1. **Communication Overhead**: As the number of processes increases, the relative cost of inter-process communication grows, particularly for the MPI-based sampling coordination.

2. **Load Imbalance**: The dynamic nature of nested sampling can create imbalanced workloads across processors, particularly when some parameter combinations require more complex equation of state evaluations.

3. **Memory Bandwidth**: The calculation of neutron star structures is memory-intensive, and memory bandwidth limitations may constrain performance at higher node counts.

4. **I/O Constraints**: At scale, file system operations (reading the crust EOS data, writing posterior samples) may introduce serialization points that limit scaling.

Conclusions and Recommendations

This scalability study demonstrates that our Bayesian inference workflow achieves significant but sub-linear scaling on the Deucalion HPC platform. Based on our findings, we recommend:

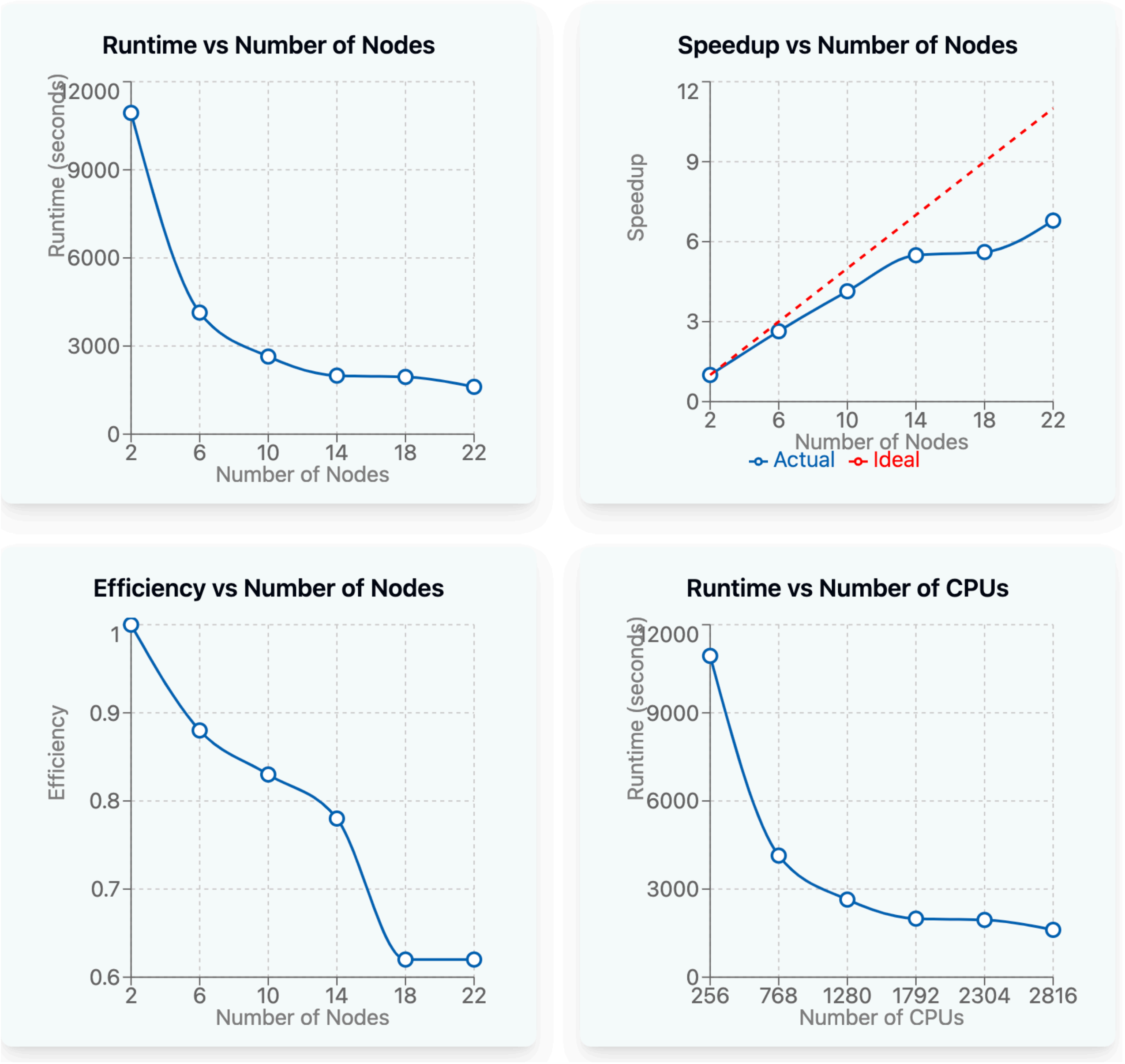
1. **Optimal Resource Allocation**: For time-critical production runs, using 14-18 nodes appears to offer the best balance between absolute performance and resource efficiency. Beyond this point, the diminishing returns may not justify the additional computational resources.

2. **Workload Optimization**: Future development should focus on reducing inter-process communication and improving load balancing to maintain better efficiency at higher node counts.

3. **Hybrid Parallelization**: Implementing a hybrid MPI+OpenMP approach could potentially improve scaling by reducing the number of MPI ranks while utilizing all available cores.

4. **Parameter Sensitivity Analysis**: Investigating which model parameters contribute most to computational load could inform targeted optimizations.

Scalability Analysis: Bayesian Inference for Neutron Star EOS



Acknowledgments

This work was supported by the EURO HPC project (EHPC-DEV-2024D12-009). The computational component of this research was carried out on the Deucalion HPC platform in Portugal, whose resources and technical assistance are gratefully acknowledged.