Phoeni6: A Container-Based Framework for Reproducible and Fair Energy Comparison of AI Workloads

Antônio Oliveira-Filho, Wellington Silva-de-Souza, Carlos A. Valderrama Sakuyama,
Samuel Xavier-de-Souza

Laboratory of Parallel Architectures for Signal Processing

RS-HPCBench-2026 @ ICS 2026
Belfast, Northern Ireland, UK
Monday, July 6, 2026

Agenda

Why This Problem Matters

- Energy consumption, unlike accuracy, is strongly hardware-dependent.
- This makes fair comparison difficult across devices, operating systems, drivers, and monitoring tools.
- As AI systems scale, transparent and reproducible energy reporting becomes a scientific and engineering requirement.

What is Phoeni6?

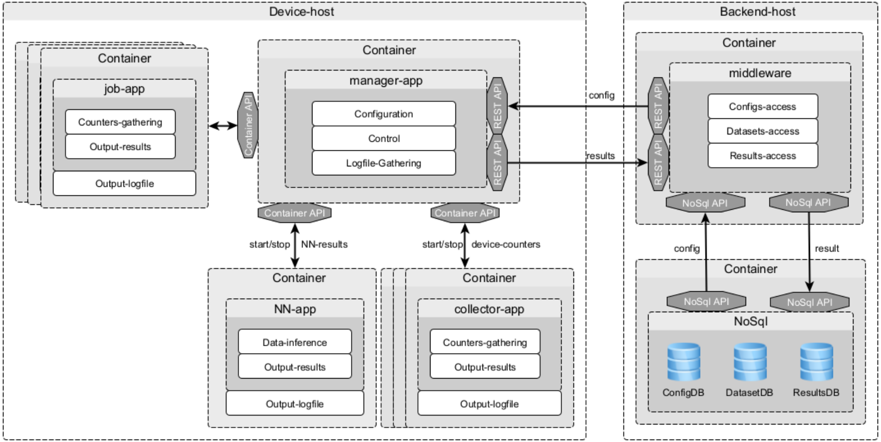
- Phoeni6 is a systematic framework for transparent energy profiling of deep neural networks.
- It is grounded in two principles:
  - **Fair Comparison (FC)** and
  - **Result Reproducibility (RR)**.
- It combines containerized tools, orchestration, and NoSQL-based persistence to automate the evaluation lifecycle.

Original paper

Phoeni6: A systematic approach for evaluating the energy consumption of neural networks
`https://doi.org/10.1016/j.suscom.2025.101172`

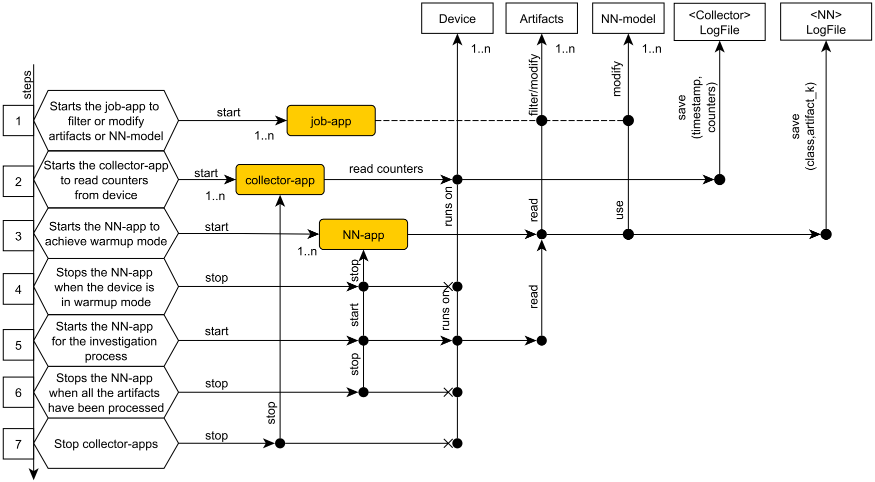
System Architecture

- Separates responsibilities between device-host and backend-host.
- Modular design supports portability, transparency, and scalability.



Execution Orchestration

- manager-app coordinates initialization, data collection, warm-up, NN execution, and logging.
- Standard flow central to FC and RR.



Main Components

- **Collector-app:** collects device counters such as power draw.
- **Manager-app:** orchestrates the full evaluation lifecycle.
- **Middleware + ArangoDB:** manage configurations, metadata, and results.
- Docker and Singularity support portability across diverse hardware architectures.

Power Measurement and Energy Calculation Methodology

- Phoeni6 measures instantaneous power using nvidia-smi
- It calculates energy by integrating instantaneous power over discrete time intervals.
- This enables granular analysis of different execution phases, such as data loading and computation.
- Energy calculated as a first-order integration of power, sample by sample

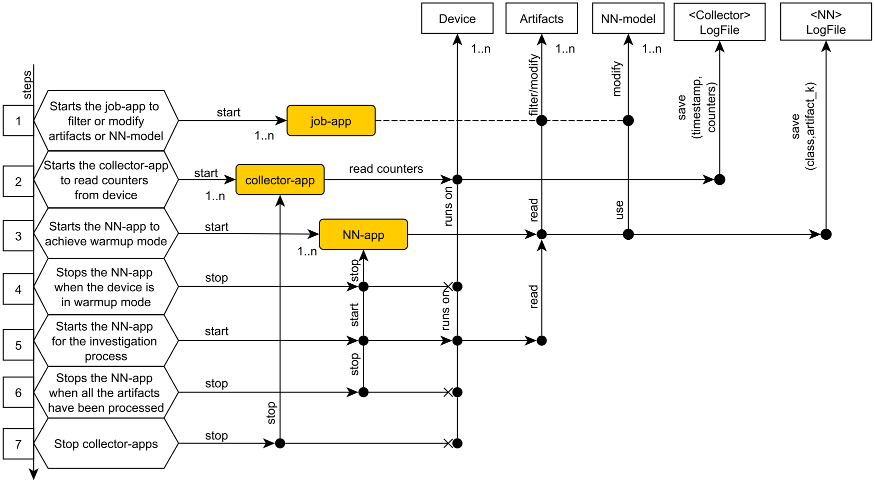
$$E = \sum_{k=1}^{n-1} P_k \times (t_{k+1} - t_k)$$

Case Study 1: Scaling Image Resizing Experiments

- Dataset: ImageNet 2012 - 23k images
- Platform: Nvidia Geforce GTX TITAN Z - 2x6 GB
- The main challenge was not only to compare energy across image sizes, but to make this evaluation feasible at scale.
- Successive image-dimension reductions can generate a massive number of derived artifacts; in the ImageNet2012 test set, creating 100 reduced versions per image would lead to about **50,000** $\times$ **100** files.
- Phoeni6 addressed this by interleaving dataset preparation and execution through dedicated **job-apps**, while collecting and storing a large volume of energy results in a structured and reproducible way.

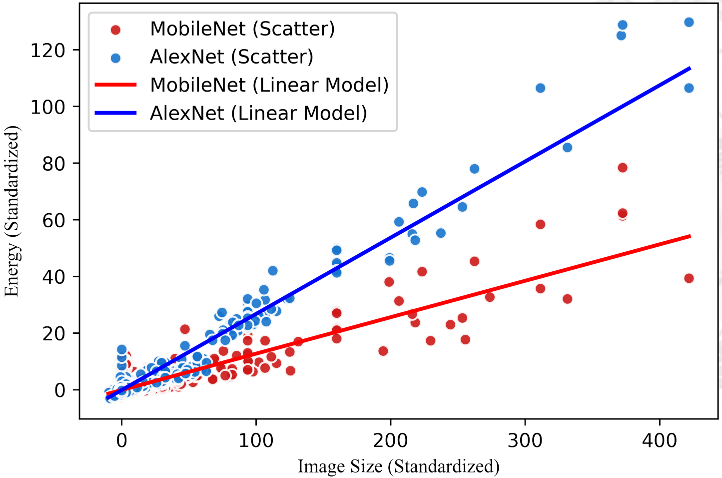
Case Study 1: Execution Flow

- Highlights modularity and adaptability during dataset resizing and energy evaluation.
- The process preserves consistency across NN models.

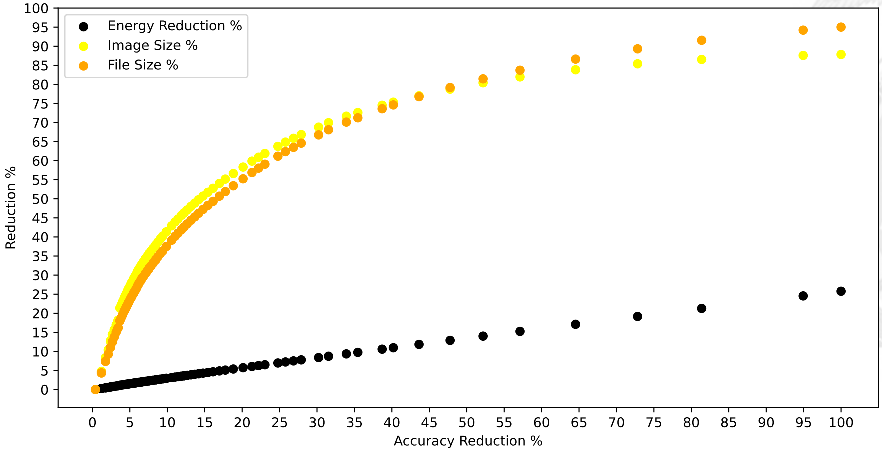


Case Study 1: Main Result

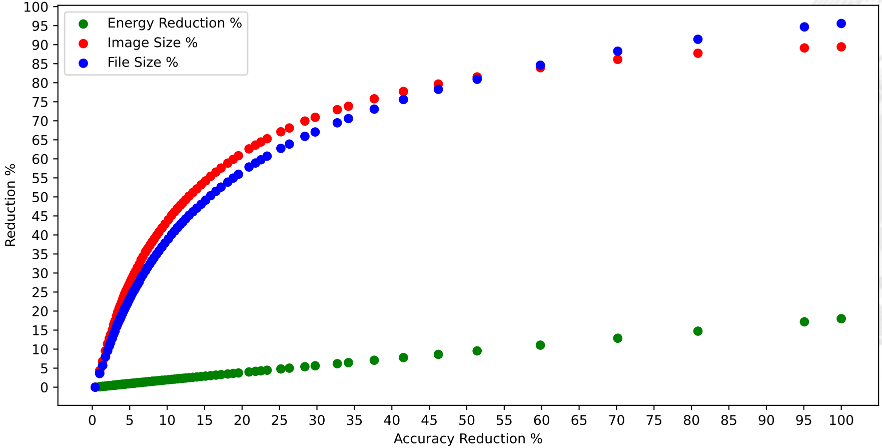
- The result indicates that model choice strongly affects energy behavior.
- In the evaluated scenario, MobileNet appears more energy efficient.



Case Study 1: Energy–Accuracy Trade-off for AlexNet



Case Study 1: Energy–Accuracy Trade-off for MobileNet

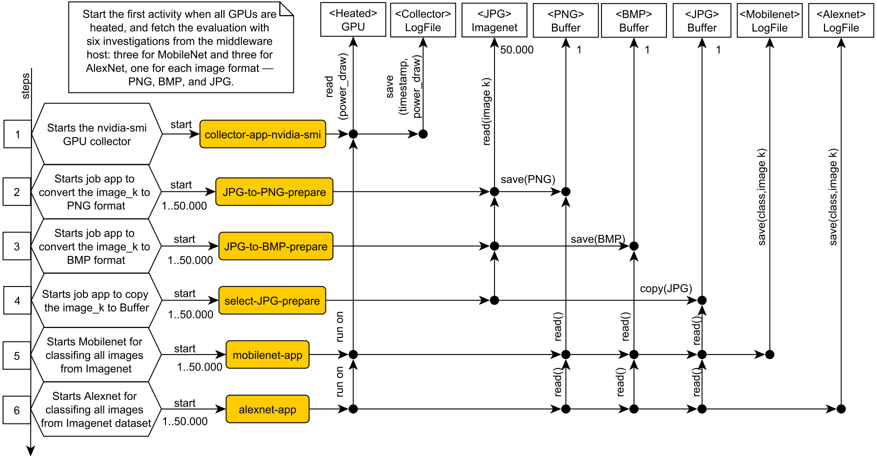


Case Study 2: Managing Format Transformation Efficiently

- Dataset: ImageNet 2012 - 23k images
- Platform: Nvidia Geforce GTX 1080 Ti - 11 GB
- The key challenge was to evaluate format-related energy costs without creating an excessive number of duplicated image sets.
- Phoeni6 handled format conversion as an intermediate **job-app** task performed during the workflow, instead of relying on permanently replicated datasets for each format.
- This approach reduced storage overhead, avoided unnecessary accumulation of repeated images, and still preserved traceability and reproducibility for the collected energy measurements.

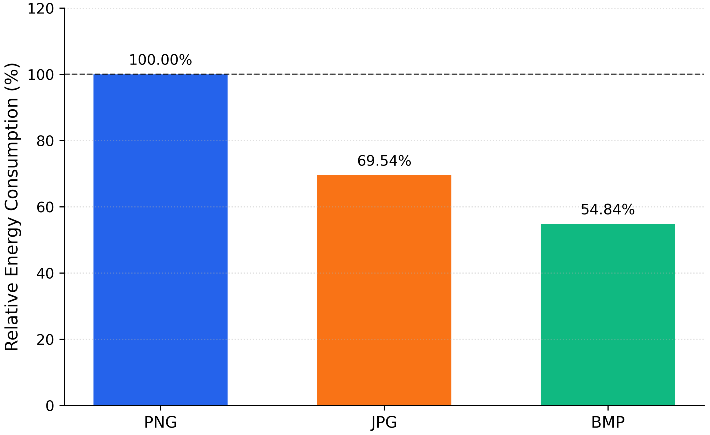
Case Study 2: Workflow Support

- Demonstrates the flexibility of Phoeni6 for varying experimental needs.



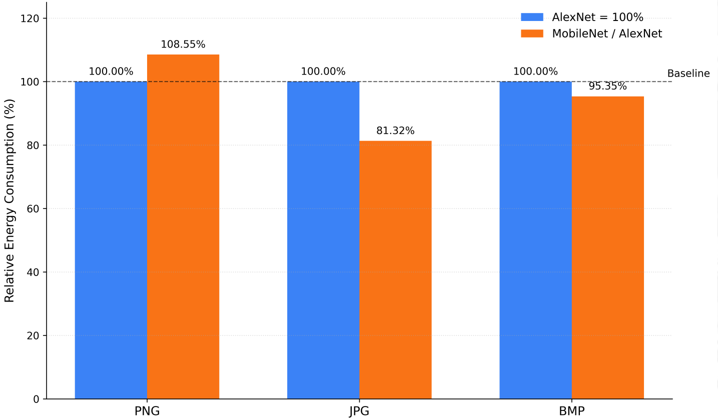
Case Study 2: Impact of Image Format on Energy

- PNG accounts for the largest usage of total energy consumption;
- BMP shows the lowest overall usage.



Case Study 2: Impact of Image Format on Energy by Network

- MobileNet better than AlexNet for JPG and BMP but not PNG.



Why This Matters for AI and HPC

- The central message from our experiments is that energy behavior changes with **model choice**, **preprocessing pipeline**, and **file format**, even under identical monitoring settings.
- This highlights the importance of **reproducible, energy-aware benchmarking** in HPC, where hardware, drivers, and tools evolve over time.
- Phoeni6's modular, container-based design provides **portability**, **traceability**, **automation**, and **comparability** for GPU workloads across devices and datasets.
- These properties are directly aligned with the RS-HPCBench agenda on reproducible and sustainable testing frameworks for HPC systems.

Conclusion

- Phoeni6 acts as a **reproducible energy benchmarking framework** for GPU-based AI workloads in HPC environments, not just as a single experiment script.
- By treating configuration and metadata as first-class data and automating warm-up, cooling checks, monitoring, and log ingestion, it reduces variability and operator error in large campaigns.
- Case studies on input size, image format, and per-class behavior show how one framework can support diverse benchmarking questions while preserving consistent measurement and analysis.
- Future work includes extending Phoeni6 to multi-node and heterogeneous clusters, integrating additional monitoring and carbon-accounting backends, and coupling energy results to schedulers such as Slurm or Flux for power-aware placement and policies.

Thank you!

MIT License: https://gitlab.com/lappsufrn/phoeni6

Questions?

RS-HPCBench-2026 @ ICS 2026