Towards Energy-Aware AI Deployment

Investigating the Interplay of Model Quantization and Hardware Platforms

Presenters:

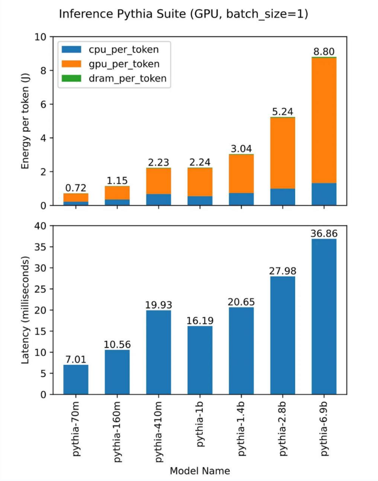
Renyuan Lu, Zinan Wang, Haoji Bian

Course:

CE 495 Energy-Aware Intelligence (EAI)

Energy Challenge & Research Motivation

LLM Energy Consumption



Training Large Transformer

1,287,000

kWh

High-Frequency Inference

25-40%

Energy Optimization Potential

Research Motivation

Core Question:

How can we achieve energy-efficient LLM deployment through the synergy of **quantization techniques**and **hardware optimization**?

Research Gap:

Existing research primarily focuses on individual factors, lacking a **systematic co-optimization framework**.

Significance

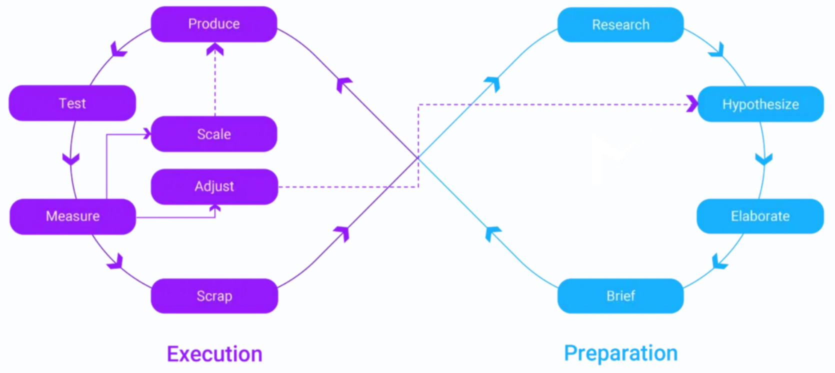
Environmental Impact:

Reducing energy consumption of AI systems is crucial for sustainable technology development.

Practical Value:

Organizations can achieve significant cost savings while maintaining model performance.

Methodology: Experimental Framework



Experimental Design Highlights

Systematic assessment of quantization impact on energy efficiency
Analysis of hardware architecture characteristics and energy consumption
Novel metrics capturing complex trade-off relationships
Validation of synergistic optimization multiplier effect

Quantization Strategy Evaluation

INT8, FP16 mixed precision, dynamic quantization
6 models with 7B parameters
5 benchmark tasks: MMLU, ARC, TruthfulQA, GSM8K, HellaSwag

Hardware Platform Analysis

Energy efficiency across 6 GPU platforms
A100 PCIE, RTX 4090, V100, etc.
Spanning Volta to Ada Lovelace architectures

Novel Efficiency Metrics

EOR = Task Performance Score / Energy Consumption (Wh)

TWEOR = Task Performance / (Energy × Inference Time)

Real-time power monitoring at 1Hz sampling

Quantization Strategy Evaluation

Quantization Strategy Comparison

Original (FP32)	Quantized (INT8) (using scale factor 16.93 and zero-point 110)
4.72	round(16.93 × 4.72 + 110) = 190
2.96	round(16.93 × 2.96 + 110) = 160
-6.48	round(16.93 × -6.48 + 110) = 0
0	round(16.93 × 0 + 110) = 110
-3.34	round(16.93 × -3.34 + 110) = 53
-5.26	round(16.93 × -5.26 + 110) = 20
8.58	round(16.93 × 8.58 + 110) = 255
2.19	round(16.93 × 2.19 + 110) = 147
-3.67	round(16.93 × -3.67 + 110) = 47

Quantization	Energy Reduction	Accuracy Loss	Best For
INT8	**25%**	0.7-0.9%	General Use
FP16 Mixed	**16%**	0.3-0.5%	High Precision
Dynamic	**10%**	0.4-0.6%	Varying Input

Key Insight:

Different quantization strategies excel in different contexts, requiring application-specific selection

Key Findings

INT8 Quantization Shows Strongest Impact:

25% energy reductionwith only 0.7-0.9 percentage point accuracy loss
DeepSeek-7B: 39.65Wh → 29.74Wh, 32% EOR improvement

FP16 Mixed Precision Provides Balance:

16% energy reduction with better accuracy preservation
Particularly suitable for high-precision requirements

Dynamic Quantization Offers Flexibility:

10% energy reduction with runtime adaptability
Ideal for varying input complexity scenarios

Performance-Energy Trade-offs

Model-Specific Variations:

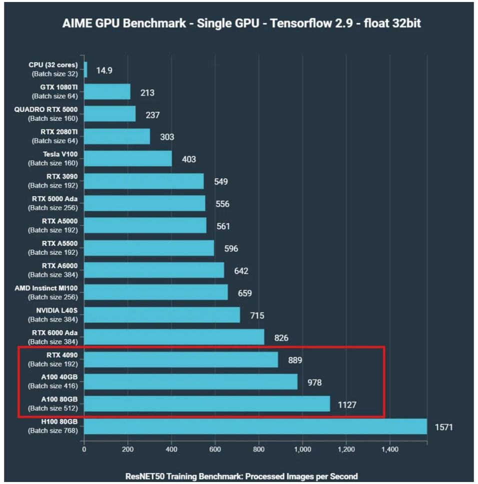
Llama2-7B showed highest quantization compatibility
Mistral-7B demonstrated best accuracy retention

Task-Dependent Effects:

Reasoning tasks (GSM8K) more sensitive to quantization
Knowledge tasks (MMLU) showed robust performance

Hardware Platform Analysis

Hardware Performance Heatmap



Platform	Architecture	Memory BW	Efficiency Gain
A100 PCIE	Ampere	1,555 GB/s	Baseline
RTX 4090	Ada Lovelace	1,008 GB/s	20-30%
RTX 4060 Ti	Ada Lovelace	288 GB/s	Best for constraints

Key Findings

A100 PCIE: All-Around Champion

Highest energy efficiency across all workloads
High memory bandwidth (1,555 GB/s) advantage
3rd-gen Tensor Core architecture benefits

Ada Lovelace Architecture: Next-Gen Benefits

RTX 4090 shows 20-30% efficiency improvements
4th-gen Tensor Cores excel in mixed precision
Better performance per computational unit

Important Discovery:

New architectures achieve 15-25% efficiency gains per computational unit

Architecture-Specific Insights

Memory Bandwidth Impact:

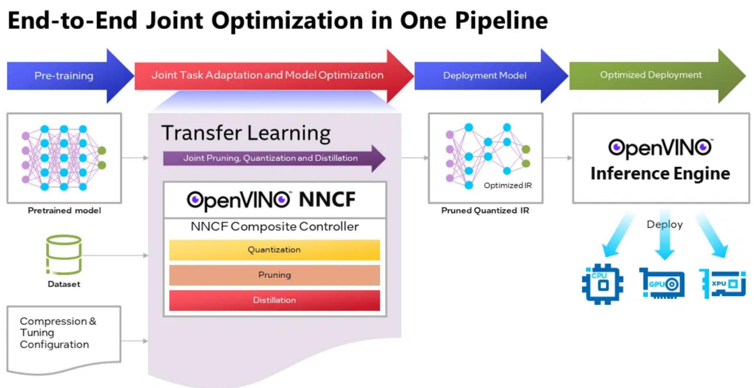
Critical factor for large model inference
Directly correlates with energy efficiency

Consumer vs. Data Center GPUs:

Consumer GPUs show surprising efficiency
Cost-performance ratio favors newer architectures

Synergistic Optimization Effect

Synergy Effect Visualization



A100 + INT8	RTX 4090 + FP16	Knowledge Distillation
40%	**35%**	**19.8%**
Efficiency Improvement	Efficiency Gain	Additional Reduction

Core Discovery:

Hardware-software co-optimization is the key pathway to energy-efficient AI deployment

Key Findings

Strategic Quantization-Hardware Matching:

A100 PCIE + INT8 quantization: **40% comprehensive efficiency improvement**
RTX 4090 + FP16 mixed precision: **35% efficiency gain**
While maintaining 98%+ baseline accuracy

Knowledge Distillation Enhancement:

DeepSeek-R1-Distill model provides additional 19.8% energy reduction
Cross-platform consistency with enhanced quantization compatibility

EOR and TWEOR Metric Validation:

Successfully capture performance characteristics missed by traditional metrics
Provide quantitative foundation for deployment decisions

Multiplicative vs. Additive Effects

Beyond Simple Addition:

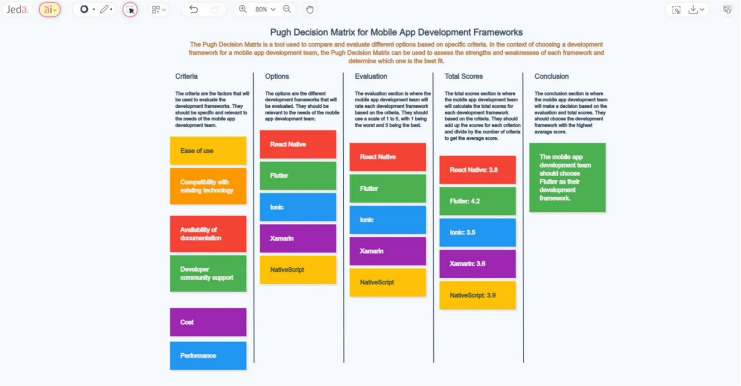
Co-optimization yields greater benefits than individual optimizations combined
Hardware-specific quantization tuning unlocks hidden efficiency potential

Architectural Compatibility:

Ada Lovelace architecture shows superior INT8 compatibility
Ampere architecture excels with dynamic quantization

Application Guidelines: Deployment Recommendations

Deployment Decision Matrix



Scenario-Specific Recommendations

Our research demonstrates that different deployment scenarios require different hardware-quantization combinations for optimal energy efficiency. Data center environments benefit from high-performance hardware with aggressive quantization strategies, while edge computing needs solutions that balance power consumption and precision.

By following these recommended configurations, organizations can significantly reduce energy consumption while maintaining model performance, contributing to sustainable AI development.

Data Center Production

Hardware:

A100 PCIE

Quantization:

INT8

Expected:

- 98% performance
- **40% efficiency improvement**

Enterprise Applications

Hardware:

RTX 4090

Quantization:

FP16 mixed precision

Expected:

- 99% performance
- **35% efficiency improvement**

Edge Computing Deployment

Hardware:

RTX 4060 Ti

Quantization:

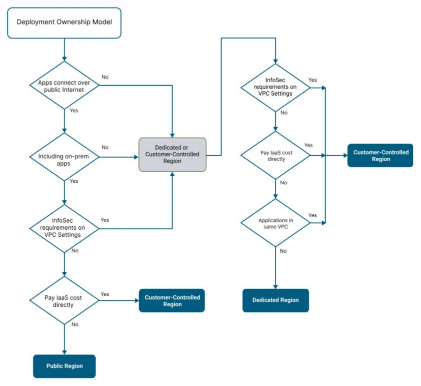
Dynamic quantization

Expected:

- 95% performance
- **25% efficiency improvement**

Decision Tree Guidance

Deployment Decision Flow



Decision Framework Value

This decision tree framework provides a systematic approach to help organizations select the optimal hardware-quantization combination based on their specific needs. By following these three key steps, deployment teams can avoid common pitfalls such as over-provisioning hardware or selecting quantization strategies unsuitable for the application scenario.

Our research shows that the right decision flow can achieve 25-40% energy savings while maintaining model performance, which is particularly important for large-scale deployments.

1
Hardware Support Assessment

- Evaluate available hardware platforms
- Determine Tensor Core support
- Analyze memory bandwidth limits

2
Application Requirements Analysis

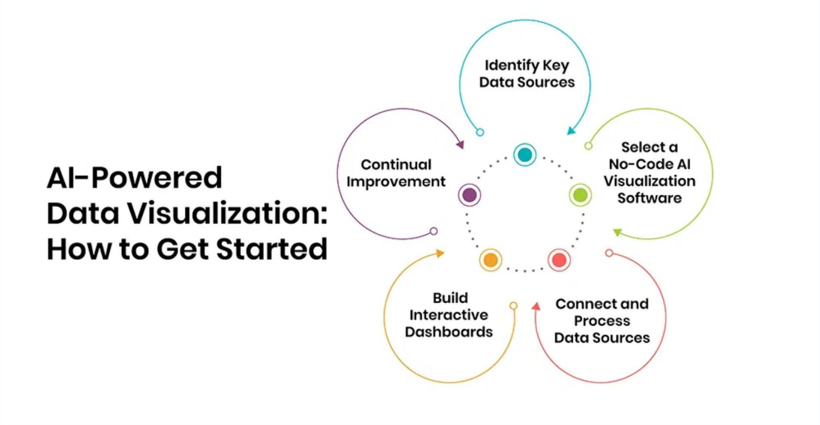
- Determine precision thresholds
- Evaluate latency tolerance
- Analyze input complexity variation

3
Cost Constraint Consideration

- Calculate total cost of ownership
- Evaluate energy cost impact
- Analyze scale deployment needs

Summary: Core Contributions

Core Contributions Visualization



Quantization Savings	Co-optimization Gain	Accuracy Retention
25%	**40%**	**98%+**

Key Findings

Quantization techniques can reduce energy consumption by 25%

Co-optimization achieves 40% efficiency improvements

Hardware-software matching is crucial

Technical Contributions

1 **First quantization-hardware co-evaluation framework**

2 **Novel EOR/TWEOR energy efficiency metrics**

3 **Evidence-driven deployment guidance framework**

Real-World Impact

As AI systems scale deployment, **energy-aware optimization** will become central to sustainable technology development.

Our work provides foundational insights and practical tools for this goal, directly applicable to production environments for significant energy savings.

Future Work & Acknowledgements

Future Research Directions

Extension to larger model scales

Exploring energy efficiency strategies for 70B+ parameter models

New hardware architecture research

Evaluating synergistic effects with next-gen AI accelerators

Real-world deployment scenario validation

Validating findings in diverse production environments

Research Impact

Our work on energy-aware AI deployment provides a foundation for sustainable AI scaling as these systems become increasingly prevalent in society.

The quantization-hardware co-optimization framework offers immediate practical benefits while opening new research directions in energy-efficient AI.

By addressing both technical optimization and deployment guidance, this research bridges the gap between theoretical advances and practical implementation.

Broader Implications

Energy-efficient AI deployment contributes to corporate sustainability goals and environmental responsibility initiatives.

Cost savings from optimized deployments can be redirected to further research and innovation.

Our framework enables more accessible AI deployment in regions with limited energy infrastructure.

Presenters:

Renyuan Lu, Zinan Wang, Haoji Bian