ConfigSpec: Profiling-Based Configuration Selection for Distributed Edge-Cloud Speculative LLM Serving

4th International Workshop on Testing Distributed Internet of Things Systems

Xiangchen Li, Saeid Ghafouri, Jiakun Fan, Babar Ali, Hans Vandierendonck, Dimitrios S. Nikolopoulos

Content

QUEEN'S UNIVERSITY BELFAST

- Need for Edge-Cloud Collaboration
- Speculative Decoding
- Configuration Challenge
- ConfigSpec
- Experimental Setup
- Results

Need for Edge-Cloud Collaboration

QUEEN'S UNIVERSITY BELFAST

Why Edge-Cloud LLM Inference?

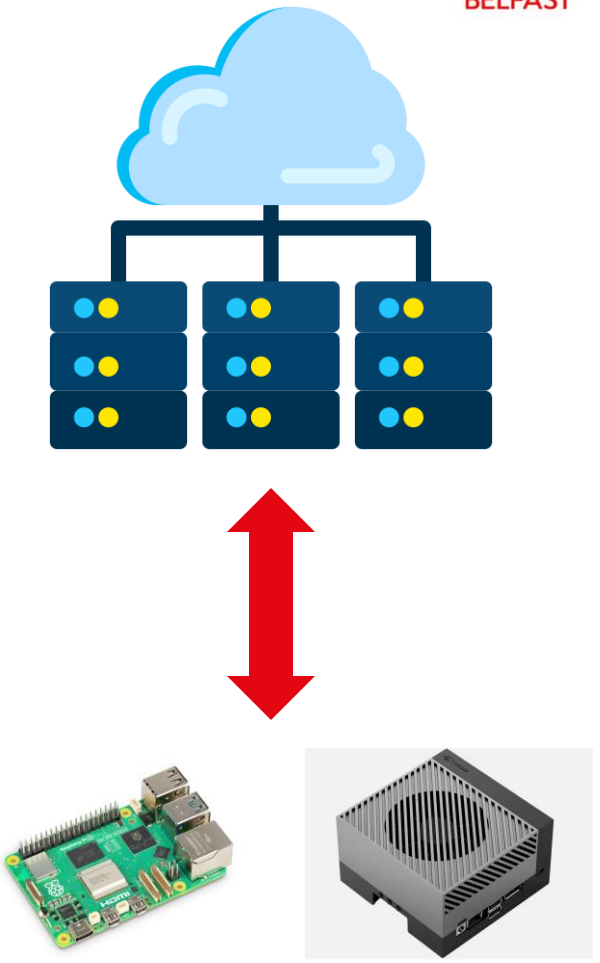
- Interactive and IoT workloads need low latency
- Data locality, privacy, bandwidth constraints
- Full LLM execution on edge is still impractical

Key Challenges:

- Heterogeneous and resource-constrained edge
- No sacrifice on correctness

Solution

- Edge and Cloud collaboration



Speculative Decoding

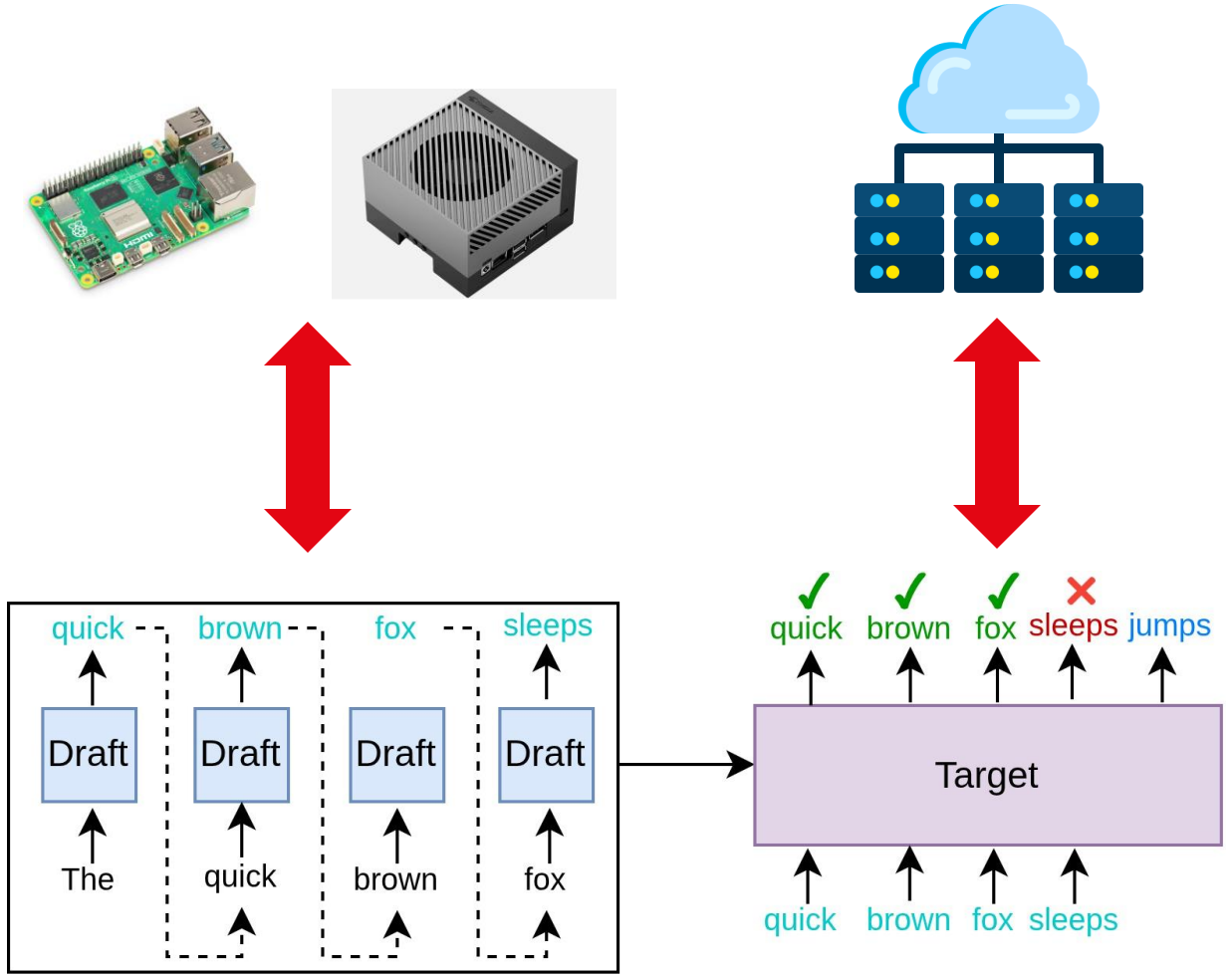
- A small draft model proposes multiple tokens
- A large target model verifies them
- Accepted prefix committed, divergence corrected

Benefits:

- Preserves target model output distribution
- Improves throughput and cost efficiency

Collaborative LLM Inference:

- Draft in Edge devices
- Target in Cloud Servers



Configuration Challenge

QUEEN'S UNIVERSITY BELFAST

Configuration Knobs:

- Draft model size & family
- Quantization level
- Speculative length
- Edge hardware pla

Reality: No single configuration dominates across all devices and objectives

Draft Model (M):

1B, 3B, 8B

Bigger ➡ higher acceptance rate but more power

Quantisation (Q):

Q4_K_M ➡ Q8_0

Lower bits conserve memory but drops quality

Speculative Length (K):

K = 2,…,10

Longer K accelerate LLM inference but increases rejection risk

Edge Device

RPi 4B, RPi 5, Jetson

Drafting speed and power substantially differ across platforms

Introducing ConfigSpec

Profiling-based configuration selection framework

Profiling:

- For each (edge device, draft model) pair measure throughput and device power P
- Measur

Goal: Enable fast, principled configuration selection before deployment

Model:

- Uses analytical models to evaluate goodput, cost, and energy

Select:

- Enumerate (M, Q, K) and rank configurations under the deployment's active objectives - latency, cost, or energy.

ConfigSpec: System Framework

Edge Devices:

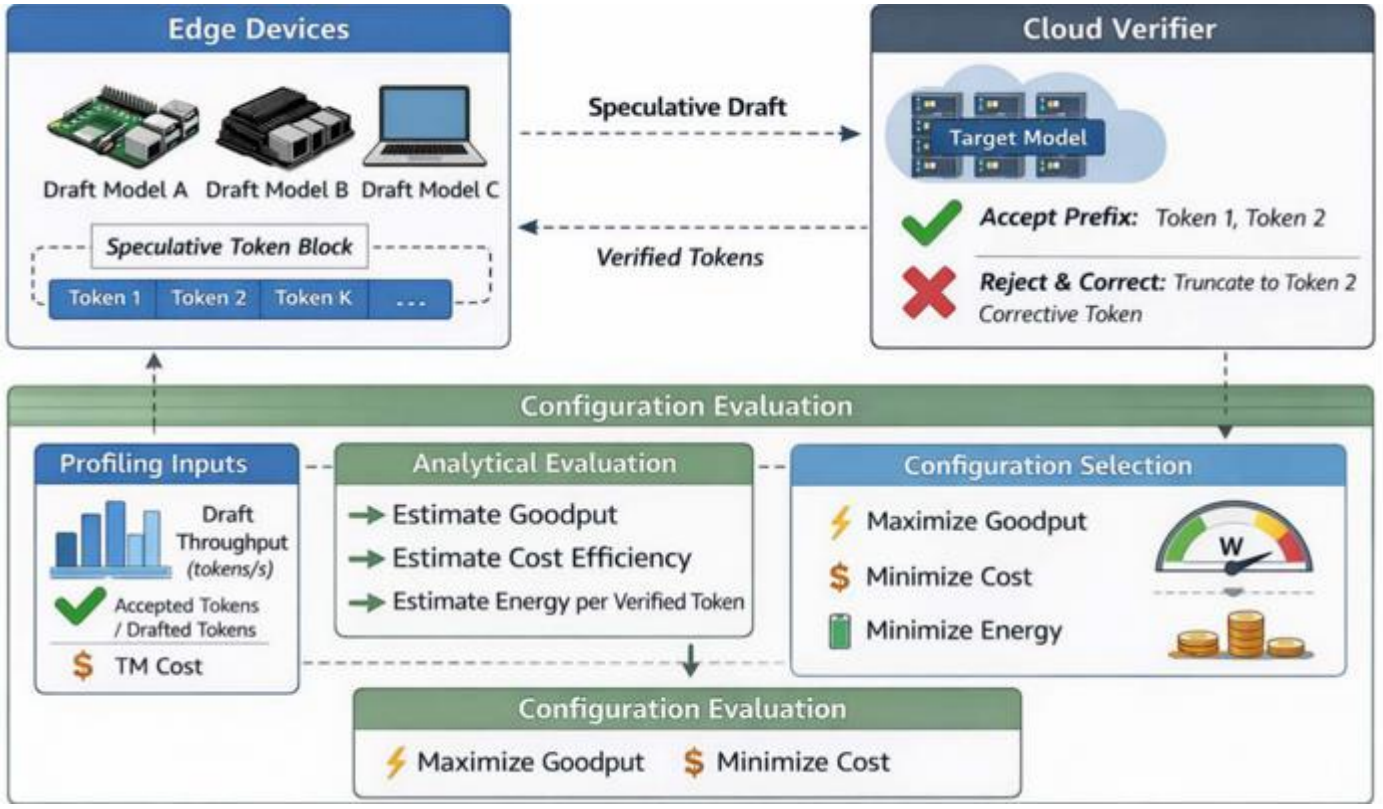
- Host draft models (e.g., Llama-3.2-1B, Qwen3-0.6B) and generate speculative tokens

Cloud Verifier:

- Hosts the target model (e.g., Llama-3.1-70B, Qwen3-32B)

Speculative Loop:

- Edge drafts K tokens ➡ Cloud verifies ➡ Cloud accepts prefix or rejects and corrects ➡ Repeat.



Analytical Modelling

Goodput:

$$G(K) = (K \cdot \alpha(K) + 1) / (K/v_d + T_{verify})$$

Tokens / second

Numerator: Expected accepted tokens plus one "bonus" token.

Denominator: Time taken (local draft time + remote verification time).

Cost Efficiency:

$$\eta_{cost} = (\alpha(K) + 1/K) / p$$

Tokens / dollar

Depends only on acceptance rate, speculative length, and cloud token price.

Energy Efficiency:

$$E = P \cdot (K / v_d) / (K \cdot \alpha(K) + 1)$$

Joules / verified token

Drafting energy divided by expected accepted tokens.

Depends on device power P and drafting speed v_d

Experimental Setup

Edge Platforms:

Raspberry Pi 4B

Cortex-A72 · 8 GB · CPU only

Raspberry Pi 5

Cortex-A76 · 8 GB · CPU only

NVIDIA Jetson AGX Orin

Ampere GPU · 64 GB unified

Models:

Drafts:

Llama-3 (1B to 8B), Qwen3 (0.6B to 8B)

Targets:

Llama-3.1-70B, Qwen3-32B

Workload:

Databricks Dolly 15K prompts

Pricing:

Llama: $0.90 / 1M tokens

Qwen: $0.59 / 1M tokens

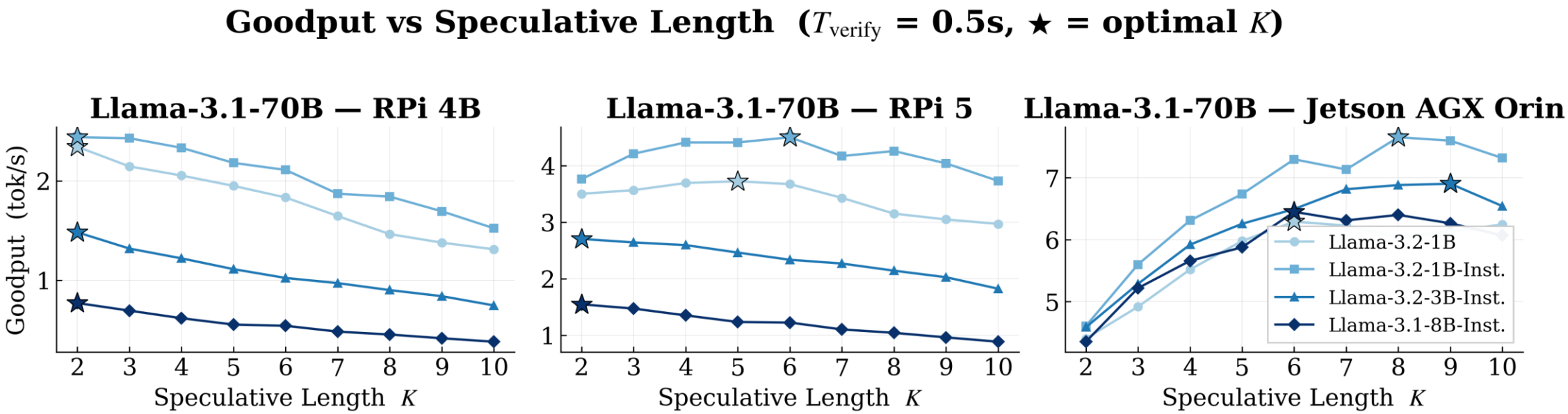
Results - Goodput Analysis

Findings:

Goodput favours the smallest, fastest draft model

Optimal K∗ is device dependent:

- RPi 4B: K∗ ≈ 2
- RPi 5: K∗ ≈ 6–7
- Jetson: K∗ ≈ 8–10 (T_{verify} bottleneck)



Results – Quantization Impact

QUEEN'S UNIVERSITY BELFAST

Findings:

6.5 tokens/sec on Jetson

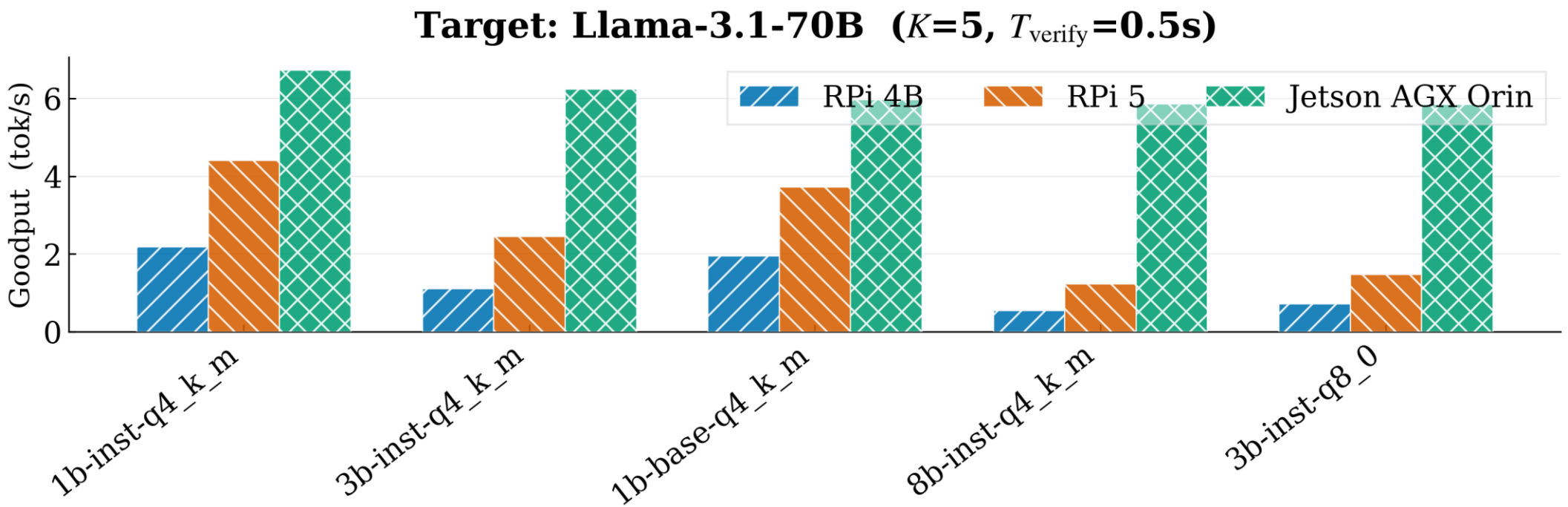
Llama 1B-Q_4_K_M

4x goodput drop for RPis

1B to 8B shift (despite α rise from 0.46 to 0.62)

1.5-2x Jetson Improvement over RPis

Verification latency bottleneck



Results – Cost Analysis

Findings:

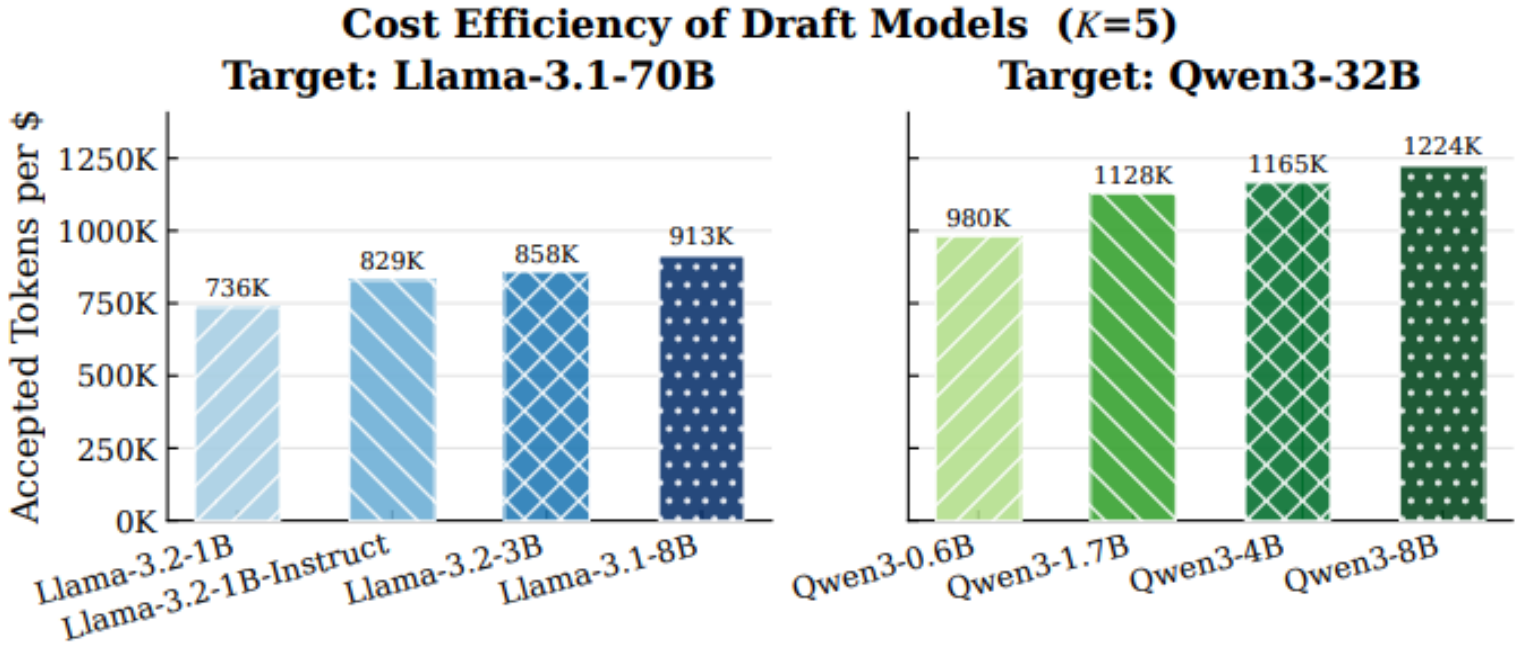
Llama 8B 19% better than the 1B

8B 25% better than 0.6B

Largest draft models win

Always maximized at K = 2

"Bonus-Token Effect" strongest at small K



Results – Energy Analysis

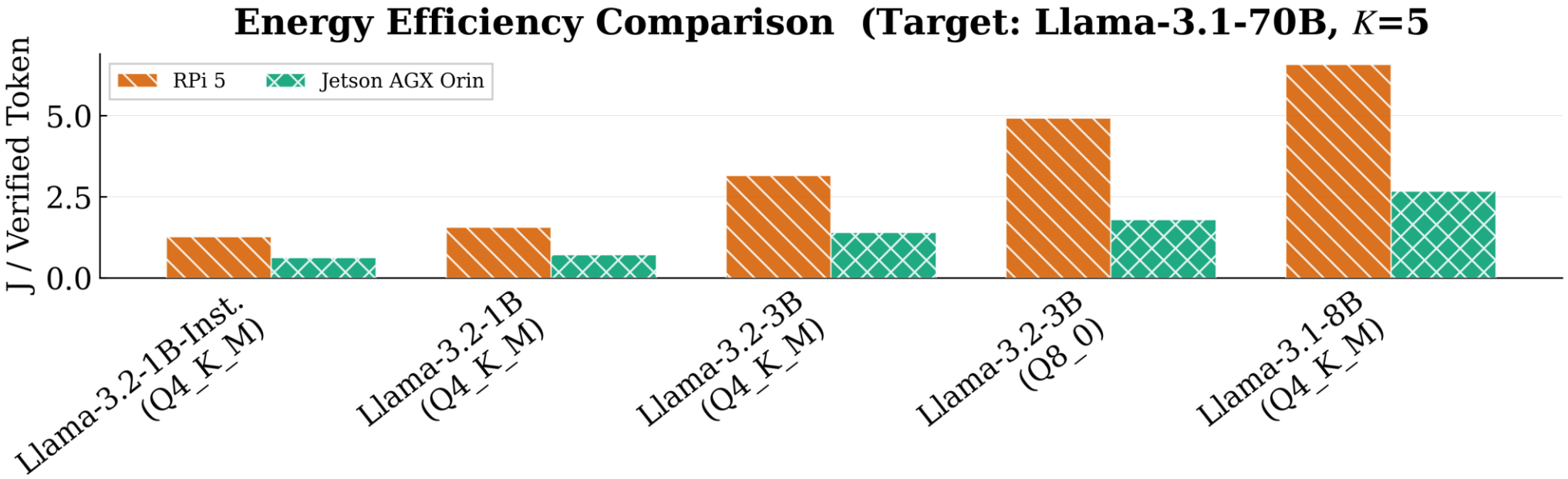
Findings:

Smallest drafts are energy-efficient

Smaller drafts minimize active compute time

Energy-optimal K=2 on all devices

Jetson outperforms RPis



Joint (M,Q,K) Analysis

Trade-Offs:

8B to 1B

- Up to 2.9× throughput
- 7.8× energy
- Cost 46%

Objective	Model Size	K
Maximise Goodput	Smallest	2-10 (Device-Dependent)

No single fixed configuration optimises goodput, cost, and energy

Conclusion

QUEEN'S UNIVERSITY BELFAST

Lessons

- Configuration choice dominates performance outcomes
- Goodput, cost, and energy impose structurally conflicting pressures
- Profiling-based selection is essential

Why ConfigSpec?

- Lightweight
- Deployment-agnostic
- Complements speculative decoding systems

Thank you