



# Multi-Scale Local Speculative Decoding for Image Generation

Elia Peruzzo   Guillaume Sautière   Amirhossein Habibian

Qualcomm AI Research



Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patents are licensed by Qualcomm Incorporated.

Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.



## Accelerate AR models (training-free)

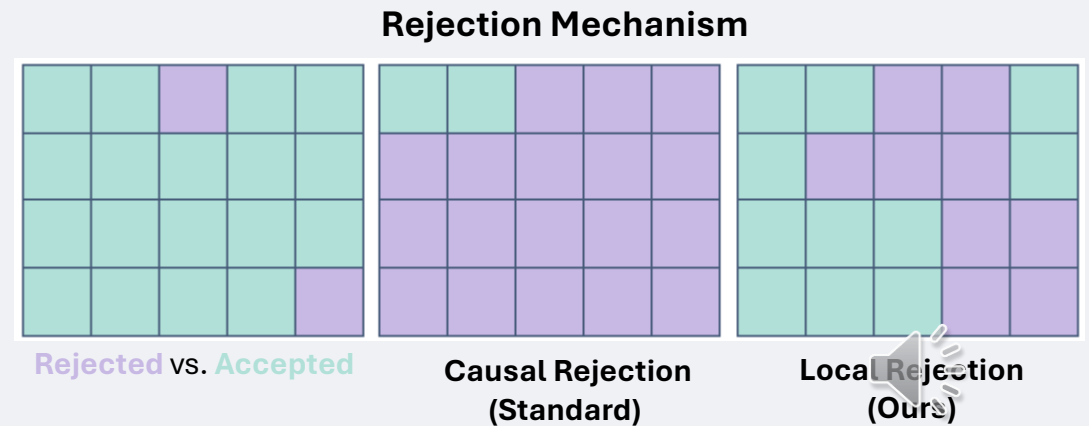
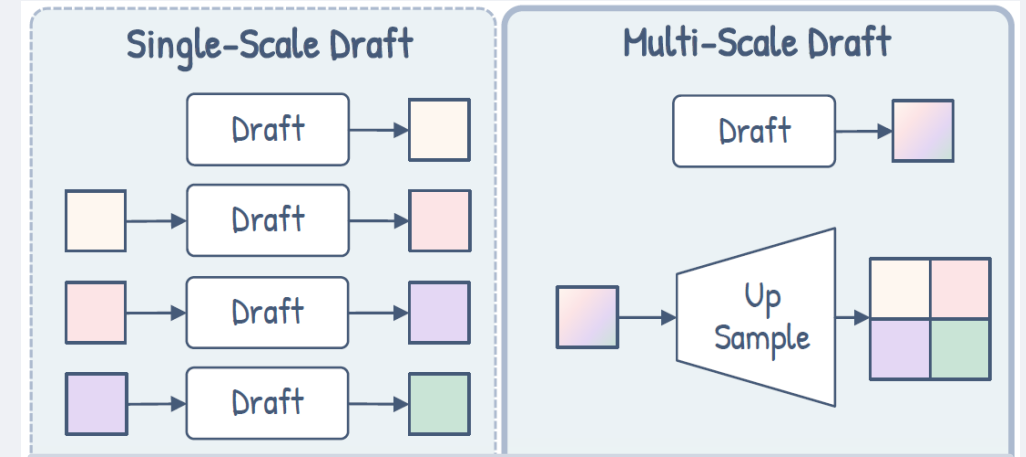
- Next Token Prediction is expensive at inference: sampling is **sequential**
  - Images are made of long token sequences (~4K for 1024p image).
  - Latency is the main blocker for interactive and real-time setting (~80s on A100).
- **Parallel Decoding**, assume (approximate) token independence and decode many tokens in parallel at inference time.
- **Speculative Decoding**, generate a draft of  $N$  tokens cheaply, then verify with the target model

The two can be combined for faster sampling, with target-model verification limits drift from the original distribution

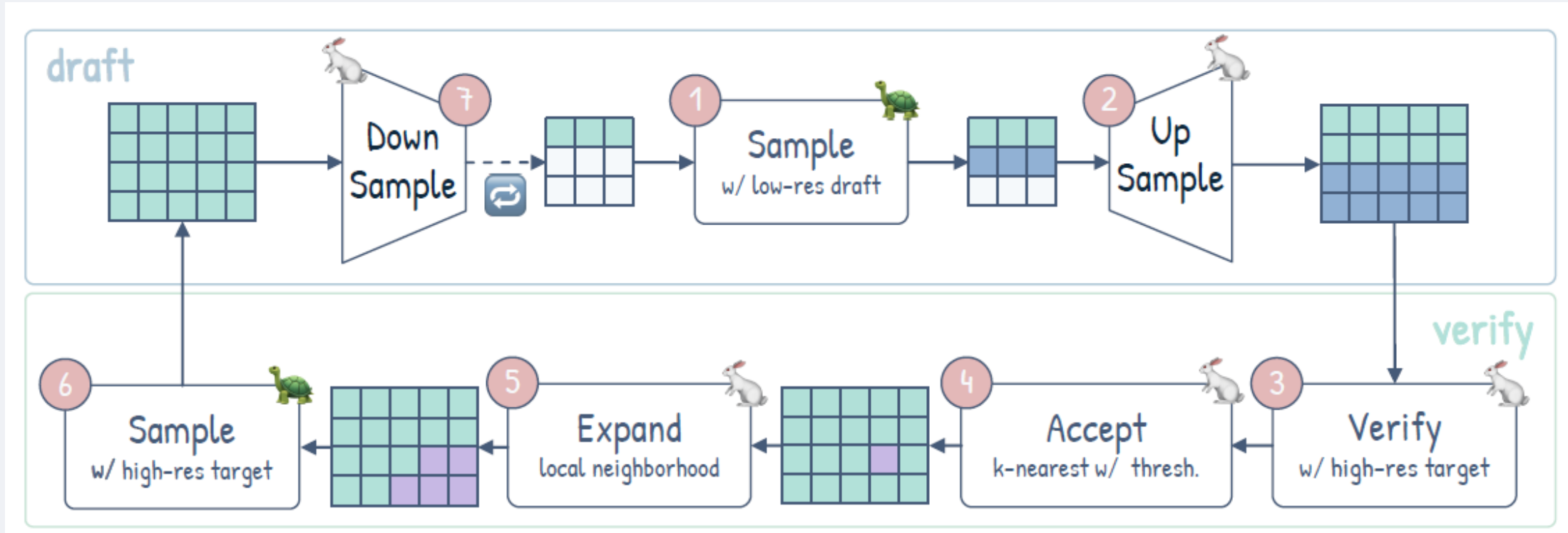


# Multi Scale Local Speculative Decoding

- Tailor SD for image synthesis to allow LLM leverage the **multi-scale inductive bias** for acceleration
- Introduce image-specific **rejection mechanism**, exploiting local bias and redundancy of visual tokens.
- We combine it with **parallel decoding** showing huge gains in latency.

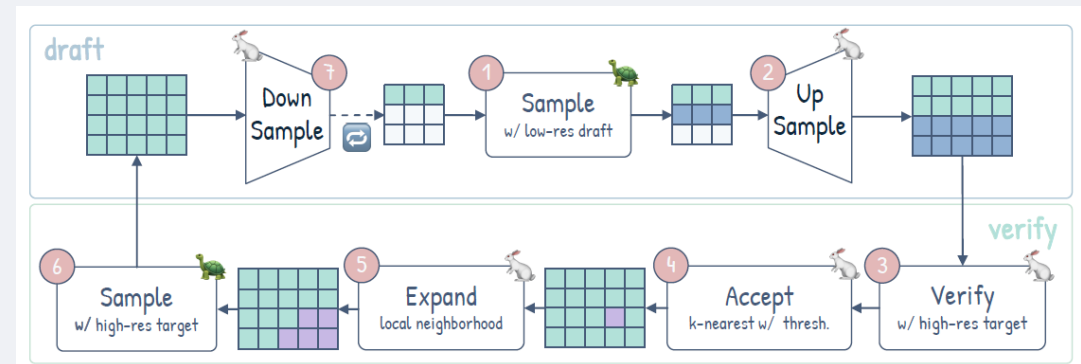


# Algorithm



# Integration with Parallel Decoding

- Sequential sampling from drafter / target still takes ~90% of the latency (Step 1 and Step 6)
- Combine speculative decoding with parallel sampling for the highest speedup:
  - Parallel sample the whole image from the **drafter**
  - Resample the rejected tokens with parallel decoding from the **target model**



# Base Model – Tar

## Vision as a Dialect: Unifying Visual Understanding and Generation via Text-Aligned Representations

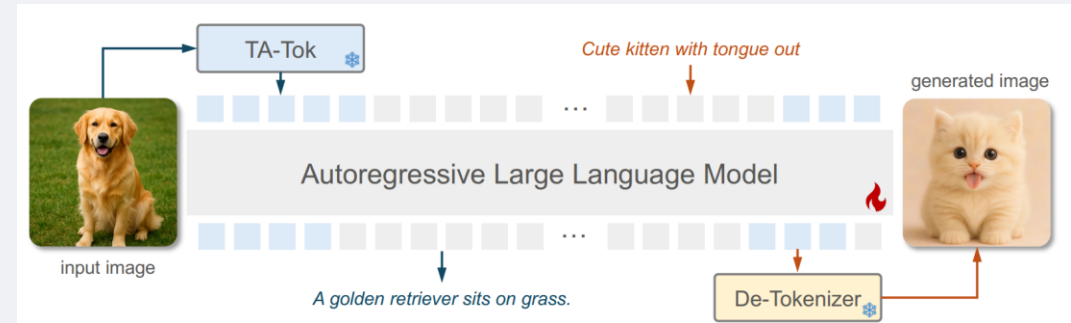
- **MLLM:**

- QwenVL-2.5 (1.5B and 7B):
- Finetuned to *generate* visual tokens (token num: 81, 169, **729**)

- **Pixel Decoding:**

- Architecture based on Llama-2 with 0.6B params
- Finetuned with the new conditioning
- Resolution specific checkpoints: 256p, 512p, 1024p.

- It is still two stage (conditioning, generation), but fully next-token-prediction.



Resolution	# Tokens	Latency [s]
256p	256	5
512p	1024	20
1024p	4096	80

Measured on A100



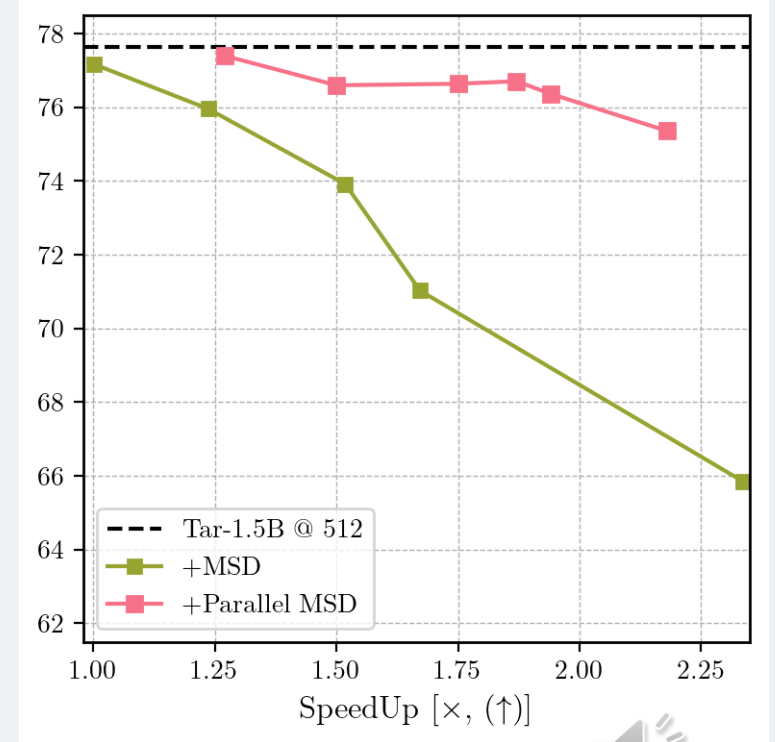
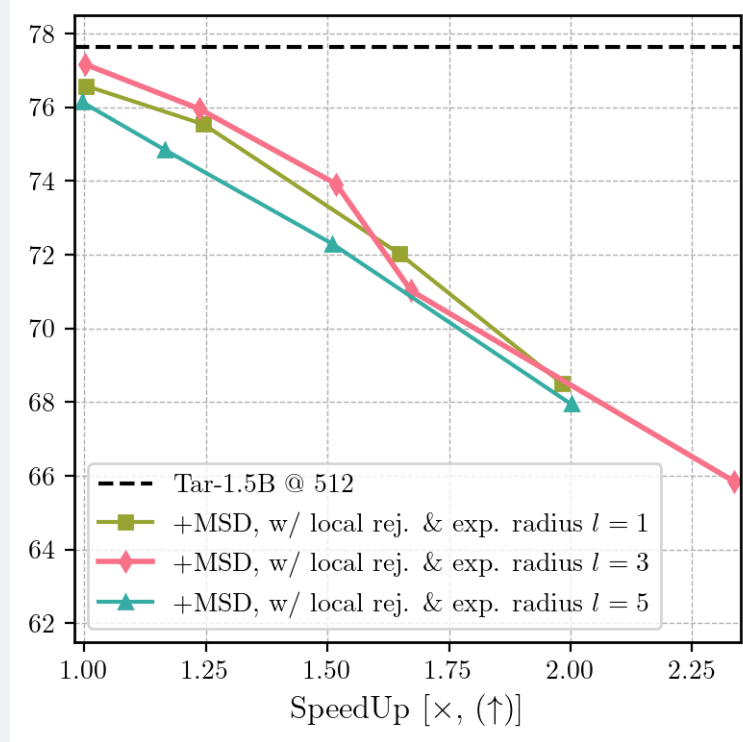
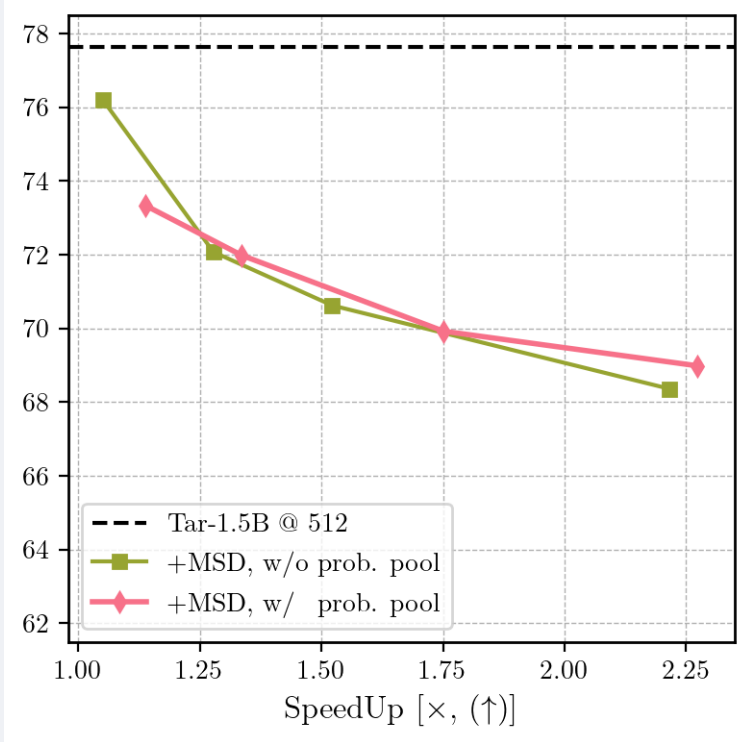
# Experiments

- **MuLo-SD** outperforms other speculative decoding competitors (**LANTERN** and **EAGLE-2**)
- When combined with parallel decoding it delivers higher **speedups up to ~5x**

	Method	Efficiency	Semantic Alignment		Perceptual Quality	
		Speedup (↑)	GenEval (↑)	DPG-Bench (↑)	FID (↓)	HPSv2 (↑)
512p	Tar-7B	1.00×	85.1	81.3	38.6	29.8 (-0.0)
	+ ZipAR-16 [16]	<u>1.88</u> ×	85.3 (+0.2)	81.0 (-0.3)	38.7 (+0.1)	29.8 (-0.0)
	+ EAGLE-2 [26]	0.76×	85.1 (-0.0)	81.3 (-0.0)	38.6 (-0.0)	29.8 (-0.0)
	+ LANTERN [20]	1.20×	84.9 (-0.2)	80.5 (-0.8)	36.9 (-1.8)	28.7 (-0.9)
	+ <b>MuLo-SD (2×</b> )	<b>2.03</b> ×	85.1 (-0.0)	80.8 (-0.5)	38.2 (-0.4)	29.5 (-0.3)
1024p	Tar-7B	1.00×	85.2	80.4	37.9	30.5
	+ ZipAR-16 [16]	<u>3.65</u> ×	85.2 (-0.0)	80.3 (-0.1)	37.9 (-0.0)	30.5 (-0.0)
	+ EAGLE-2 [26]	0.83×	85.2 (-0.0)	80.4 (-0.0)	37.9 (-0.0)	30.5 (-0.0)
	+ LANTERN [20]	1.45×	82.9 (-2.3)	80.5 (+0.1)	34.6 (-3.3)	29.4 (-0.9)
	+ <b>MuLo-SD (4×</b> )	<b>5.33</b> ×	85.4 (+0.2)	80.8 (+0.4)	34.8 (-3.1)	29.5 (-0.8)

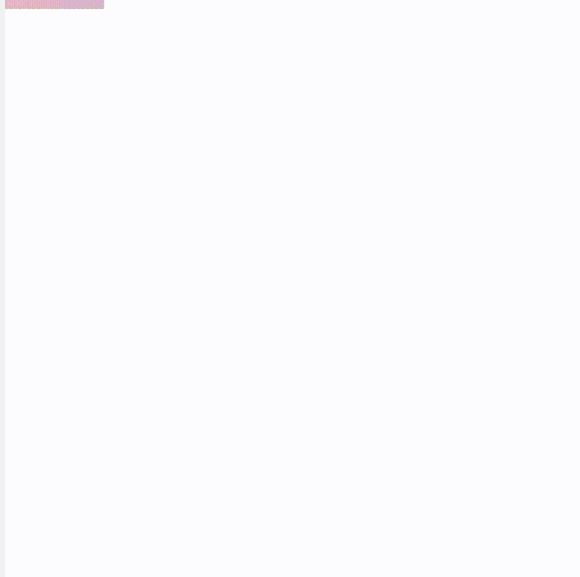


# Experiments

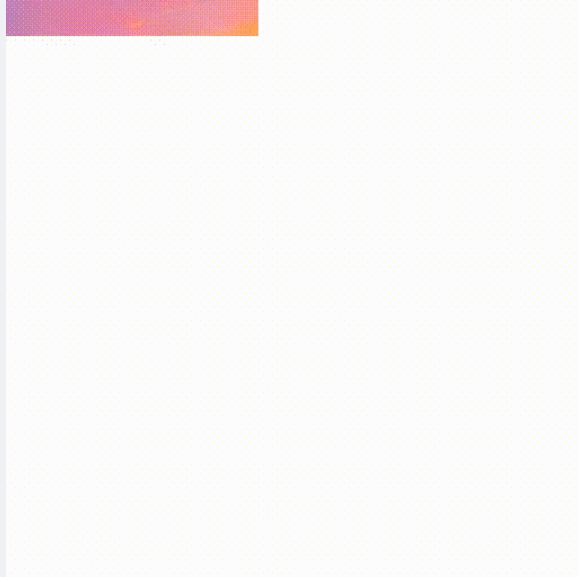


# Demo

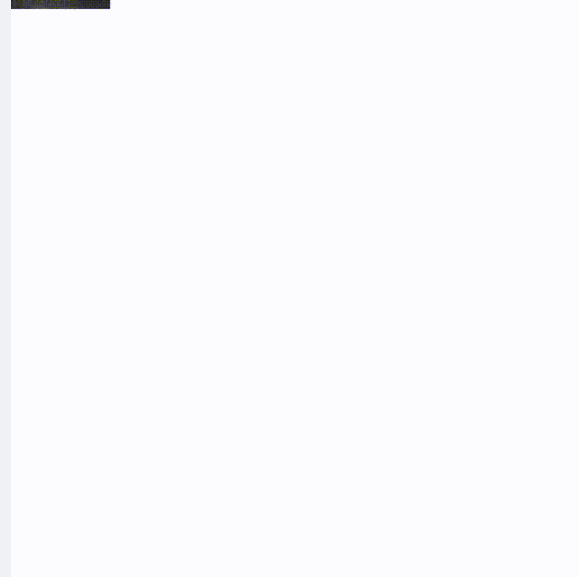
Tar @ 1024p



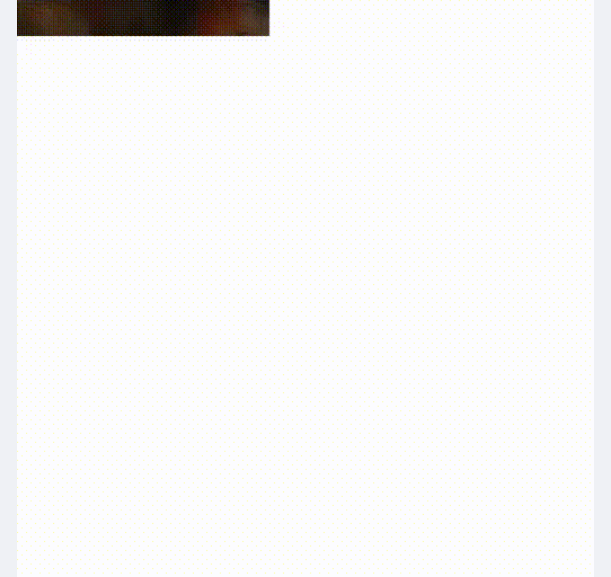
Tar @ 1024p + **MuLo**



Tar @ 1024p



Tar @ 1024p + **MuLo**



# Thank you

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

© Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm and Snapdragon are trademarks or registered trademarks of Qualcomm Incorporated. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patents are licensed by Qualcomm Incorporated.

Follow us on: [in](#) [X](#) [@](#) [v](#) [f](#)

For more information, visit us at [qualcomm.com](http://qualcomm.com) & [qualcomm.com/blog](http://qualcomm.com/blog)

