Quality-Aware Prompt Scheduling for Efficient Text-to-Image Inference Serving Shubham Agarwal*, Saud Iqbal, Subrata Mitra (Adobe Research, India)



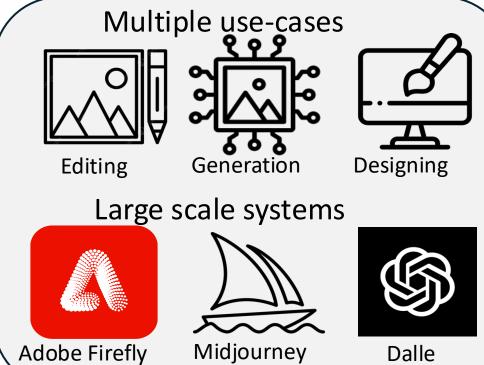
* pre-doctoral researcher

Challenges and Motivation

Approximate-Caching^[1] accelerates generation Text-to-image services Use-case Challenges: (Intermediate Image Prompt: A hyperrealistic image of a Norwegian Hound Noise for a similar prompt **Overall Goals** 1. Real-Time Needs vs. Generation **Time**-Intensive Process Quality 2.High Costs of GPU Prompt: Lion with t Infrastructure Uses intermediate noises from past hyper realistic System generations for *skipping* initial steps 3. Load Fluctuations with Throughput High Peak Loads Same model runs at Similarity with past prompt SLO decides approximation level different approximation Violations \checkmark Technical Challenges: Loads completely Prompt-to-model affinity is unknown and not used different model variants Scaling High Load Overheads due Designing Generation Costs to Model Switching Accuracy Scaling^[2] • Quality is determined by More Less Prompt-Model Synergy S [1] Shubham Agarwal et al. Approximate caching for Accurate Accurate efficiently serving diffusion models. (NSDI 2024) • Getting more Servers is [2] Sohaib Ahmad et al. Proteus: A high-throughput Horizontal Scaling Costly and Unpredictable inference-serving system with accuracy scaling.

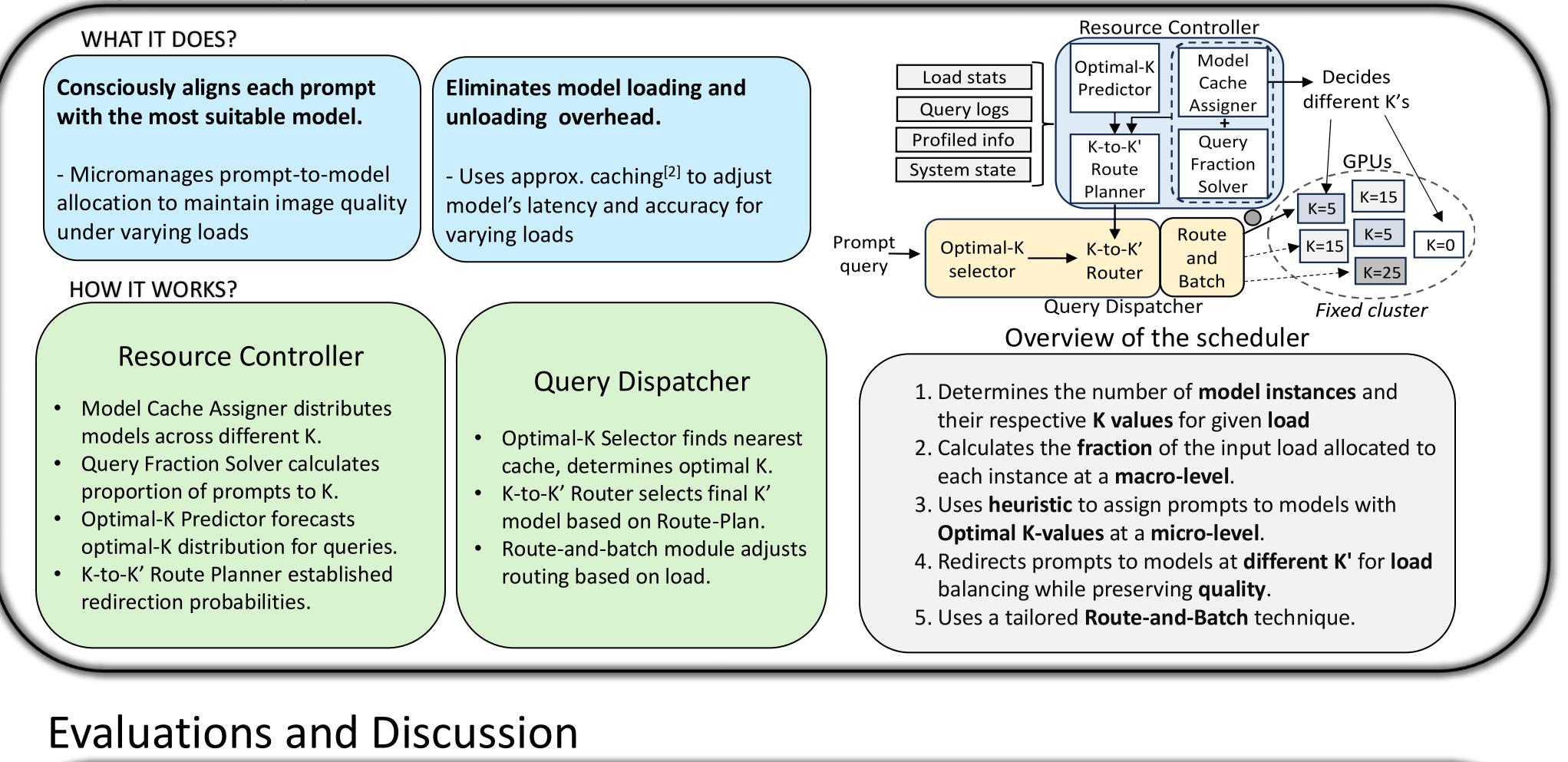


Diffusion: Iterative Denoising Process



[3] Zijie J Wang et al. Diffusiondb: A large-scale

Proposed Approach



 CLIPPER-HA[4] ••••• NIRVANA[2] --- Our System CLIPPER-HT[4]Proteus[3]

High quality

Low violations

100 200 300 400 500 600 700 800

100 200 300 400 500 600 700 800

Time (in min)

Performance of System on

(%)

Quality(in

Relative

100

96

93

90

87

84

210

60

30

0

viol

SLO

CLIPPER-HA[4] CLIPPER-HT[4] NIRVANA[2] PROTEUS[3]

5 130

ർപ്പ¹²⁰

0 110

년 100

Quality

Avg

۲<u>۲</u> 0.20

0.15

94

89

OURSYSTEM

နည်း Evaluation set-up

- SD-XL models with steps $K \in \{0, 5, 10, 15, 20, 25\}.$
- Cluster of 8 A100 GPUs.
- Twitter production and synthetic workloads traces.
- Real user prompts from • DiffusionDB^[3].

Results

Baselines face SLO violations and quality loss due to prompt-agnostic routing and loading overheads.

Maintains >90% quality and <5% SLO violations with prompt-aware approx. variant selection.

Twitter trace workload Performance Achieves up to 10% higher quality, 40% higher throughput, and 10X lower latency SLO-violations.

Aggregated System



- 1. Improves the generation quality, even under high load.
- 2. Uses a novel algorithm to optimally match the prompts to approx. cache model variants.
- 3. Future focus: Expanding to other model families.
- 4. Leveraging heterogeneous serving environments.