ML Infrastructure @ Tencent Hy

Full-stack infrastructure for rapid model iteration

Speaker: Guangming Sheng
RL Infra Lead @ Tencent HY

Pretraining Team

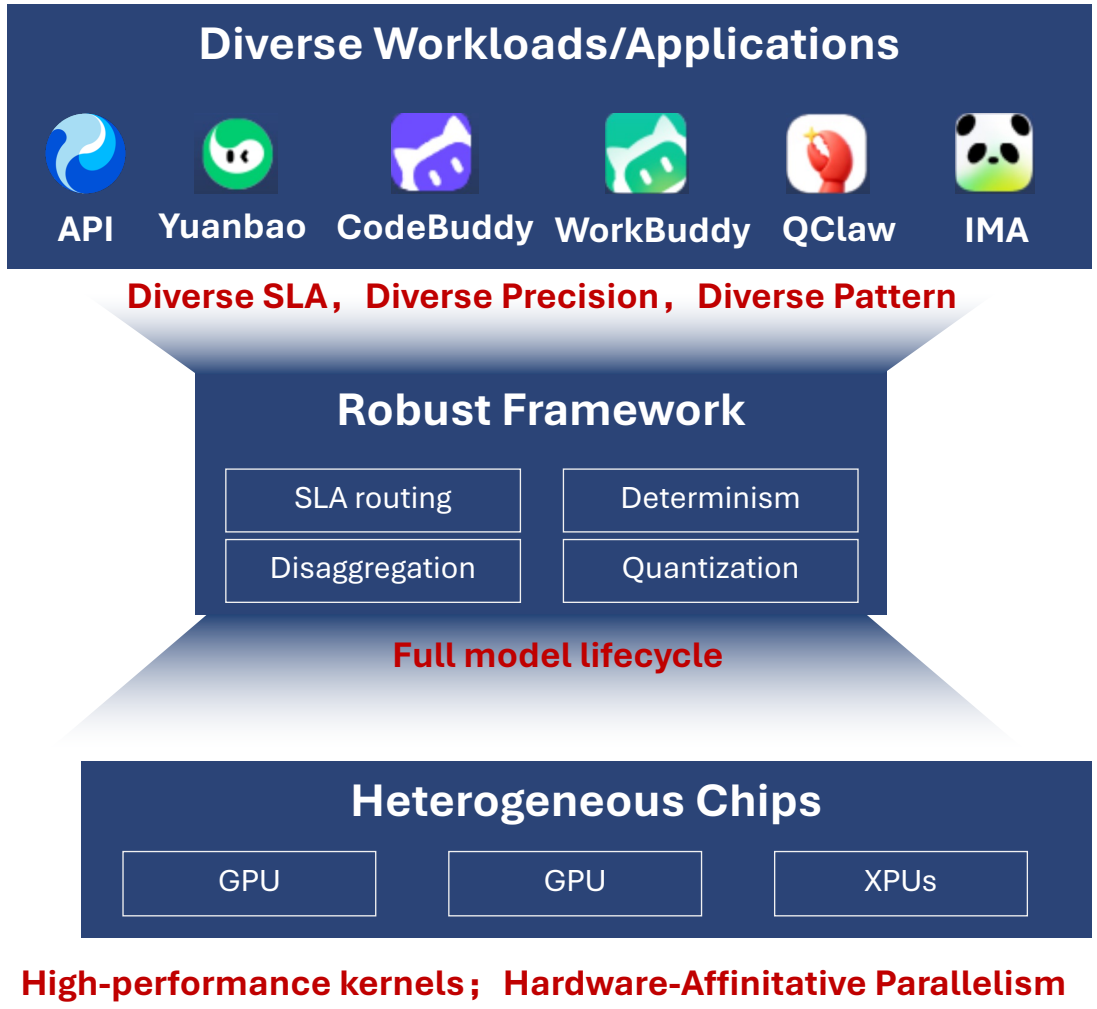
- Squeeze MFU in all modalities
- Robust and Accurate

Inference Team

- Meet diverse SLA
- Super-optimize heterogeneous chips

RL Post-Training Team

- Training on any Products
- System-Algorithm co-design



Pretraining / Inference

Training reliably at frontier scale, serving efficiently under diverse product SLAs

Tencent Hy

Pretraining: infra as research discipline

Resilience & trust

Tens-of-thousands GPU fault tolerance, SDC detection, full-stack telemetry, and checkpoint/restart policies.

Squeezing every FLOP

Custom attention / MoE / all-to-all kernels plus auto-parallel search across TP, PP, EP, SP and CP.

New training shapes

Million-token contexts, native multimodal batches, low precision FP8→FP4/MX, and agent-driven ops.

Inference: diversity is the hard case

Diverse SLA × precision × request pattern × hardware

SLA routing	Disaggregation	Quantization	KVCache
attention kernels	MoE / GEMM	GPU / XPU	Eval - Serving

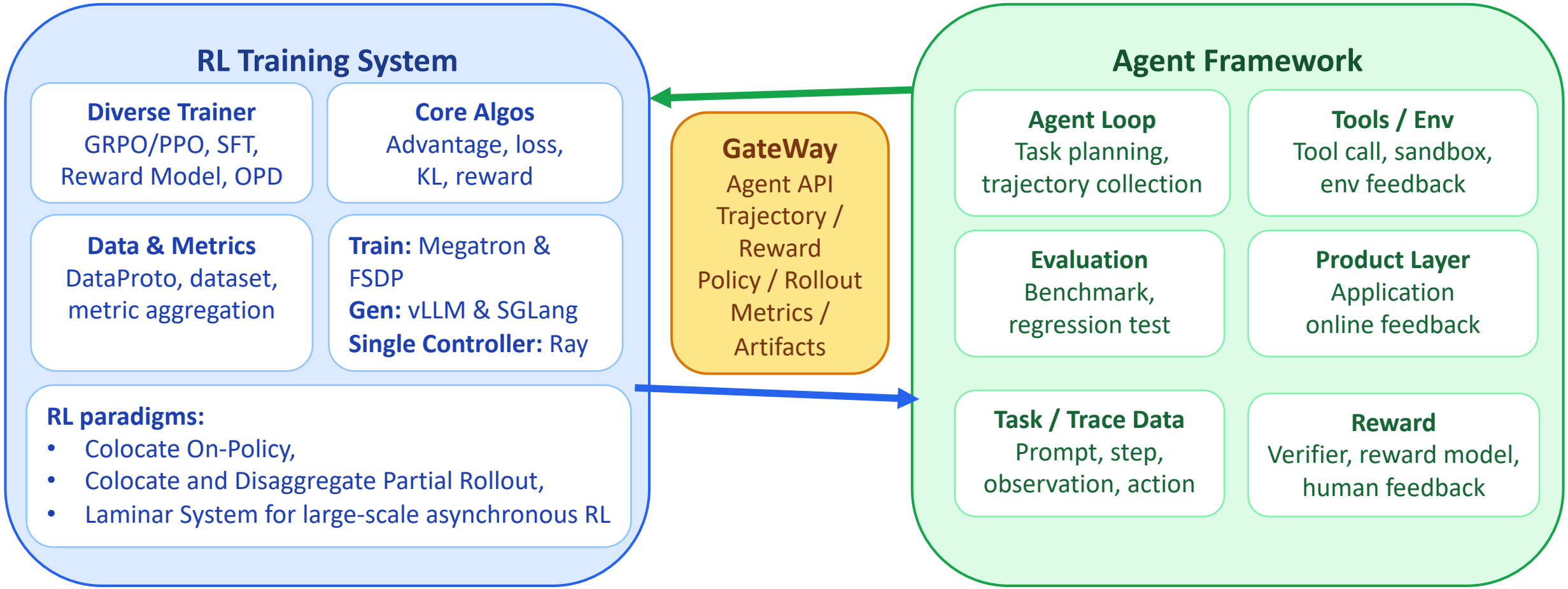
+30%
Throughput on HY3-preview via HPC Ops

3.7×
TTFT reduction via kernel co-design

https://github.com/Tencent/hpc-ops

RL Post-Training: Bridging the Train & Inference

Unified RL/RM/OPD training, agent rollout, evaluation, and product feedback



Contact: petersheng@tencent.com

Recruitment Inquiries: txqy@tencent.com