Tencent HunYuan: Building a High-Performance Inference Engine for Large Models Leveraging NVIDIA TensorRT-LLM

Yifu Sun
Meng Wang
2025.3.19

Outline Overview

Innovate. Invest. Advance: Tencent's Hunyuan AI Model Journey

Commit to be Longtermism in Technology
Bridging Utility and Engineering Breakthroughs

2020-2022
Launch 100B Ad. Rec AI Model
Launch Trillion Hunyuan LLM
Leading the Way on Prestigious Leaderboards & Pioneering Algorithm Innovation

2023.05
Enterprise-Grade AI application Integration
Empowered **Core Business Verticals** (Tencent Ads, Meetings, Docs) via **Hunyuan AI One-Stop** Business Fine-Tuning Platform

2023.09
HunYuan LLM Released to The Public
Hunyuan AI One-Stop Business Fine-Tuning Platform Fully Open to pulic

2024.1
HunYuan Deliver High-Performance MOE model
Breaking through the scale and performance boundaries
Tackle Chanllenges of costs and resources

2024.5
HunyuanDiT T2I Model Opensource
Powerful Multi-Resolution Diffusion Transformer with Fine Grained Chinese Understanding

2024.6
Release HunYuan Multimodal MOE Model
No.1 in China on the August SuperCLUE-V Chinese Multimodal Model Benchmark

2024.9
HunyuanVideo Opensource
A Systematic Framework For Large Video Generation

2024.11
Hunyuan3D 1.0 Opensource
Unified Framework for Text to 3D &Image to3D Gen

2025. 2
Release Hunyuan Next-Gen LLM Model Turbo-S
The MoE Frontier Redefined: HunYuan Turbo-S with Mamba-Transformer Core Define Next-Gen Efficiency

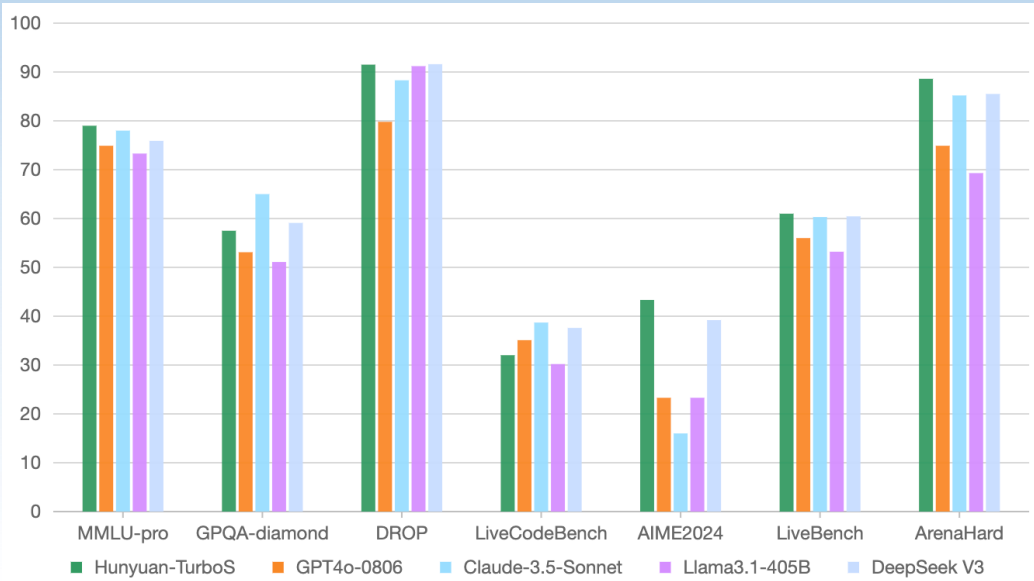
Trillion-MoE Frontier : Tencent Delivers Top-Tier AI Model

Next-Gen MoE
Tencent HunYuan Tubor-S

Top-Tier Performance

Think Faster

Reply Smarter



		Hunyuan-TurboS	GPT4o-0806	Claude-3.5 Sonnet-1022	Llama3.1-405B	DeepSeek V3
Knowledge	MMLU	**89.5**	88.7	88.3	88.6	88.5
	MMLU-pro	**79.0**	74.9	78.0	73.3	75.9
	GPQA-diamond	57.5	53.1	**65.0**	51.1	59.1
	SimpleQA	22.8	**38.2**	28.4	17.1	24.9
	Chinese-SimpleQA	**70.8**	59.3	51.3	50.4	68.0
Reasoning	BBH	92.2	91.7	**92.6**	89.2	92.3
	DROP	91.5	79.8	88.3	91.2	**91.6**
	ZebraLogic	**46.0**	31.7	35.1	30.1	38.5
Math	MATH	**89.7**	75.9	78.3	73.8	87.8
	AIME2024	**43.3**	23.3	16.0	23.3	39.2
Code	HumanEval	91.0	90.0	**95.0**	89.0	89.0
	LiveCodeBench	32.0	35.1	**38.7**	30.2	37.6
Chinese	C-Eval	**90.9**	76.0	80.0	72.7	86.5
	CMMLU	**90.8**	77.3	81.2	75.4	83.5
Alignment	LiveBench	**61.0**	56.0	60.3	53.2	60.5
	ArenaHard	**88.6**	74.9	85.2	69.3	85.5
	IF-Eval	88.6	85.7	**89.3**	86.0	86.1

Faster, Stronger, and More Efficient

More experts

Smaller Activation Volume

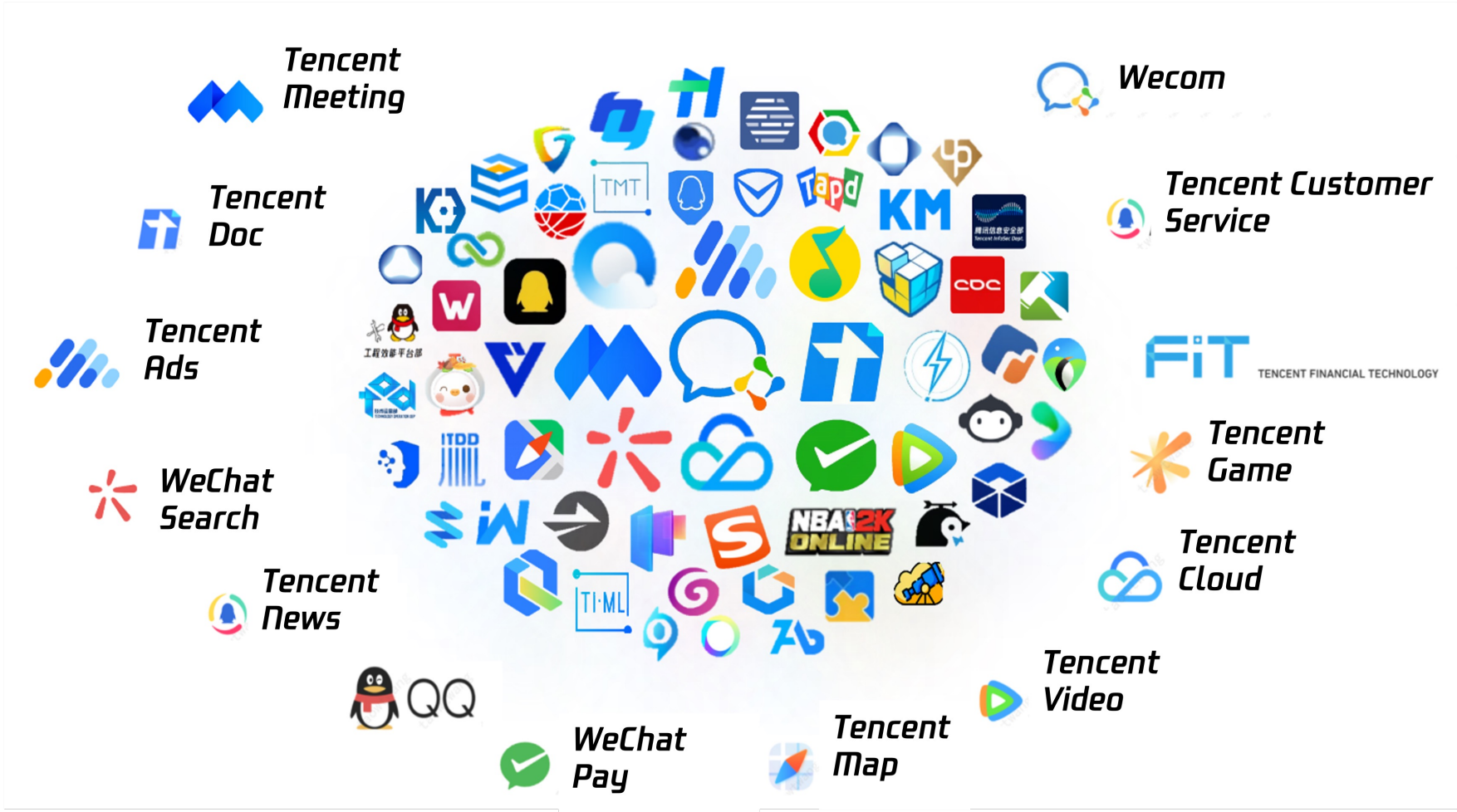
Training cost **50%** ↓

Inference Cost **70%** ↓

Decoding Speed **100%** ↑

70% cheaper than Tencent Hunyuan Turbo

From Labs to Real-World: Hunyuan LLM Now Drives 50+ Tencent Business Verticals



Over 700+ Business Applications

Pioneering R&D, Empowering Reality: Tencent Hunyuan AI Model in Application

--Empowering Multimodal Gen-AI Application Inference

Applications

Business	Tencent products Advertising, WeChat, browser, League of Legends, Tencent Meeting	Tencent Cloud API/Public Cloud Finetuning
One-Stop PaaS	OpenAPI&Business Refinement	

Tencent Hunyuan Gen-AI Model Family

Domain Specific Model	Finance	Game	Film/Anime	Code	Ads	...	
Model	Multimodal Understanding of LLM	Multimodal Generation of LLM	Text to Text	Text to Image	Text to Video	Text to 3D	LLM-TTS /ASR

Tencent Angel Machine Learning Framework

Engineering platform

- 4D Parallel
- Linear Acceleration
- Model Compression
- Kernel Opt
- PTM Training Framework
- HCF Inference Framework

AngelHCF:Hunyuan High-Performance Inference Framework

Application	Text-to-Text	Multimodal understanding Text-to-Text	Multimodal generation Text-to-Image/Video	voice LLM-TTS/ASR	Multimodal

AngelHCF

--Full Stack Engineered Excellence, Outperforming Open-Source AI Frameworks by 130%

Service	Triton Server				
Scheduling	Inflight Batching	Chunked Prefilled	Parallel decoding	KVCache/Prefix Cache	PD separation
Compression	Quantification (8/4 bits)	Distillation	Sparsification	Prune	Activation Cache
Software Acceleration	Video Memory Management Optimization	Low precision calculation W8A8C8/W4A8C4	Kernel custom optimization CLA/Mamba	Graph Optimization	Compile Optimization

Communication Library	NCCL/TCCL communication

Outline Overview

Introduction to Tencent's HunYuan Model and Taiji Angel HCF Framework

Optimization of inference in HunYuan large language models

Optimization of HunYuan Diffusion Model inference

Summary

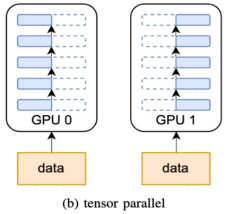
HunYuan MoE Model Inference Acceleration - Parallel Strategy

Challenge

Tensor Parallel ?

☑ More Computational Parallism

✕ Heavy All-Reduce Communication

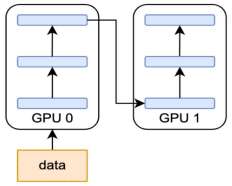


(b) tensor parallel

Pipeline Parallel ?

☑ Low Communication Overhead

✕ Relay on Micro Batch Overlap

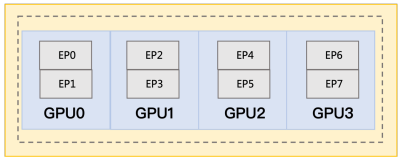


(c) pipeline parallel

Expert Parallel ?

☑ Performance Advantages on Non-NVLink GPU

✕ Hard to Balance Load



Solution

3D Hybrid Parallel = TP x PP x EP

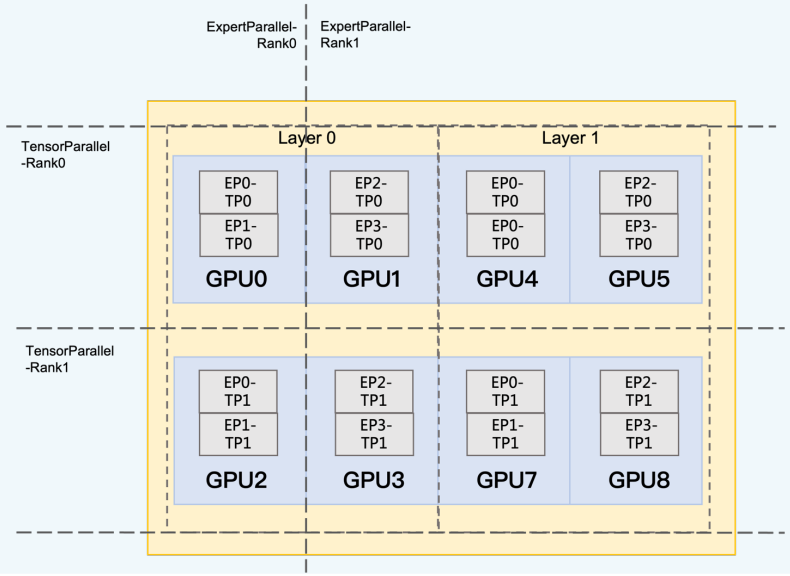
1. **All-Reduce Optimization in TP: 11% Latency Drop**

   ---Optimizatize the communication latency in NCCL

2. **Micro Batch Scheduling in PP: 20% Throughput**

   ---Eliminate Overlap Bubbles

3. **Load Balance Optimization in EP: >10% Latency Drop**

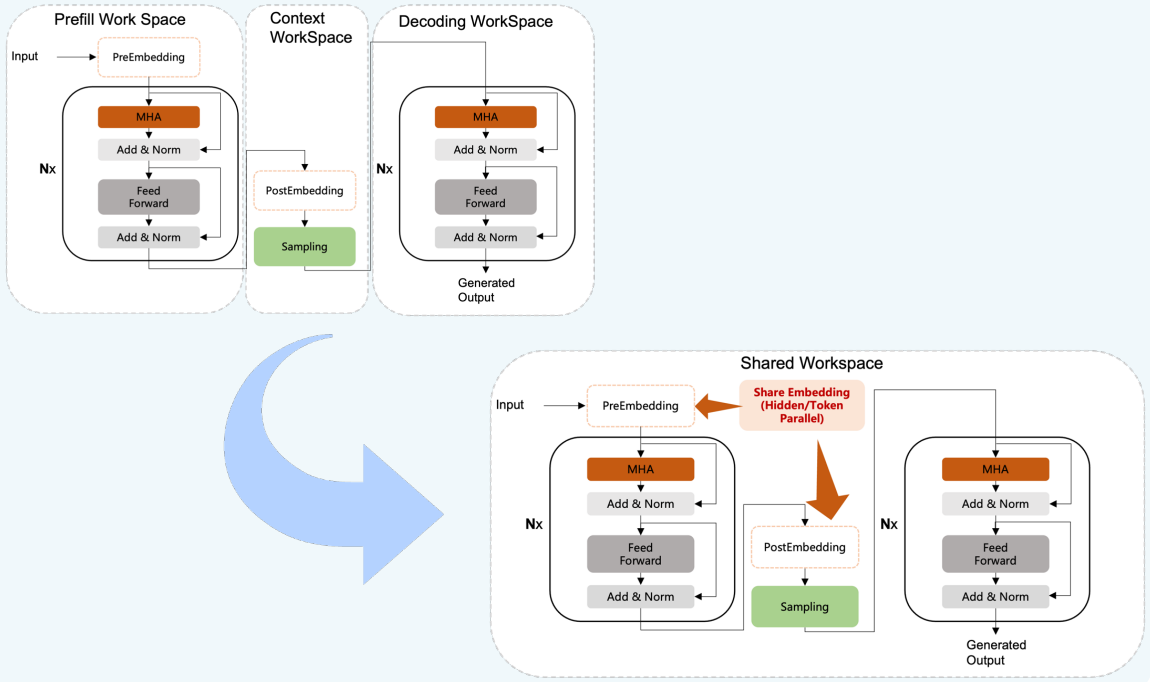


HunYuan MoE Model Inference Acceleration – GPU Memory Optimization

Technical Approch 1:Optimize Redundent Memory

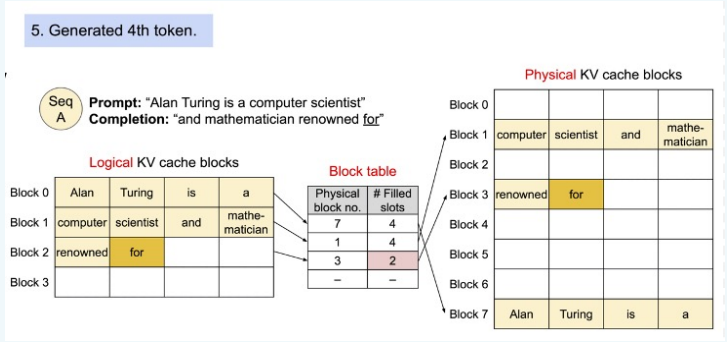
- Share Workspace of Prefill/Sampling/Decoding
- Share PreEmbedding and PostEmbedding
- Use Hidden/Token Parallel in Embedding

Tradition: 37.6 GB → -20.75% → Shared: 29.8 GB → -25.50% → Parallel: 22.2 GB

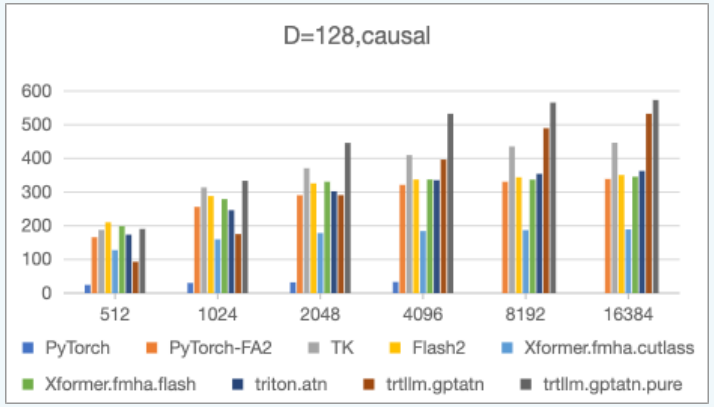


Technical Approch 2: Optimize Memory of Attention

- PagedAttention: BatchSize ↑ GPU Memory:55% ↓ Invalid Decoding ↓



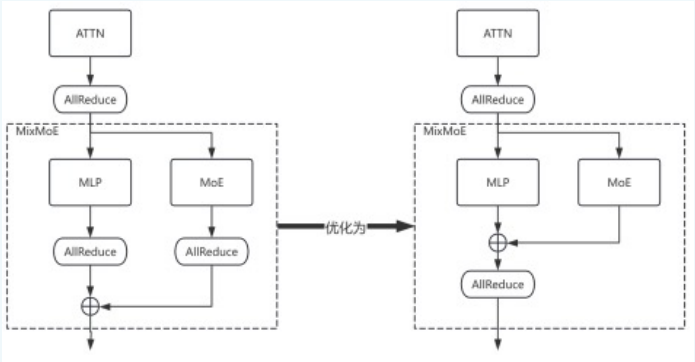
- Attention Acceleration with multi-block decoding is significant when sequence length > 1k.



HunYuan MoE Model Inference Acceleration – Communication and KV Cache Optimization

Solution1:Optimize Communication

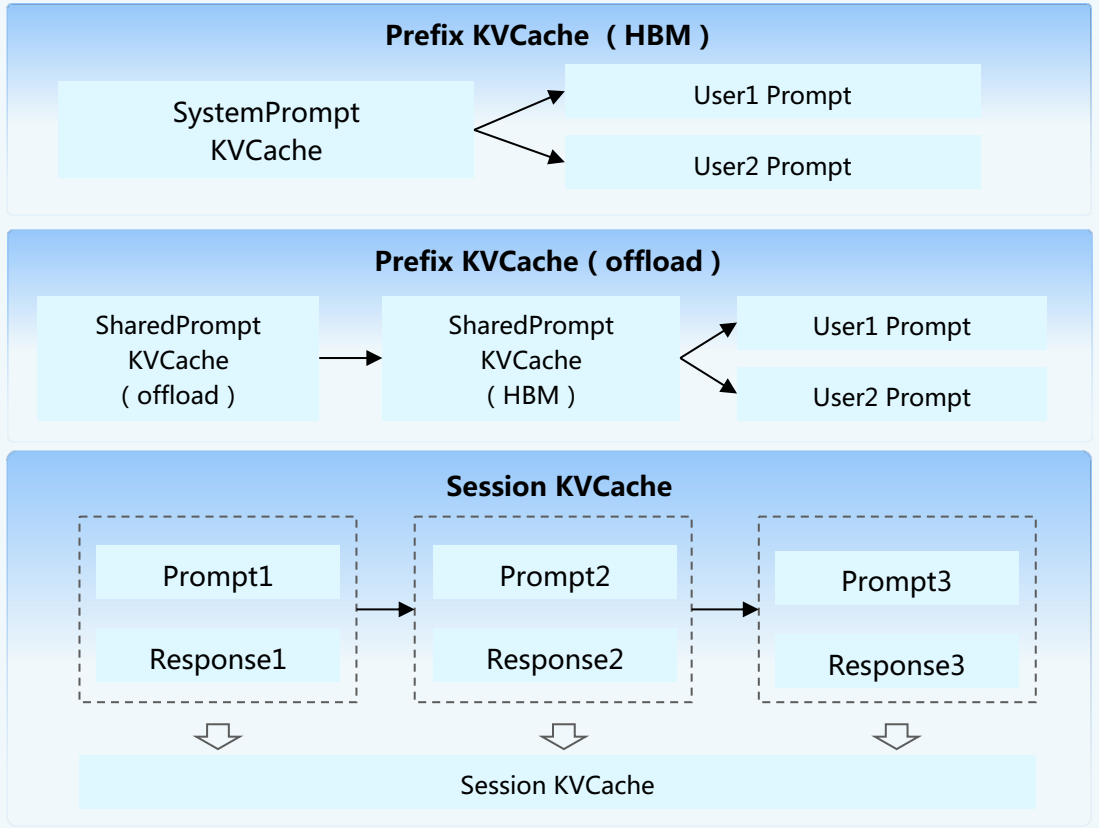
- **FFN Communication Optimization 30%:**
  - First Token Lantency Reduce **14.6%**
  - End2End Inference Time Reduce **9.7%**
- **Customized All Reduce**
  - End2End Inference Time Reduce **10%**



Benifits

- Comprehensive Applied to All Tencent Hunyuan Powered Bussiness Vertical
- Tencent Hunyuan-Turbo: First Token Latency Reduce **20%**, Throughout Increase **30%**
- Tencent Hunyuan-Role-Play:First Token Latency Reduce **40%**, Throughout Increase **50%**

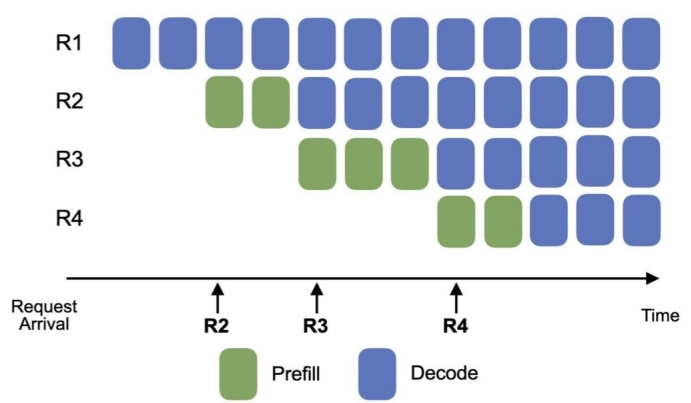
Solution2: Optimize KV Cache



HunYuan MoE Model Inference Acceleration – Optimization of Long Seq Inference

Solution1: Chunked Prefill

- Divide the input into fixed-length chunks and process KV cache Sequentially.
- Compute KV cache for each chunk incrementally while caching intermediate results.
- Parallelization can be applied to overlap computation and data transfer, maximizing GPU utilization.

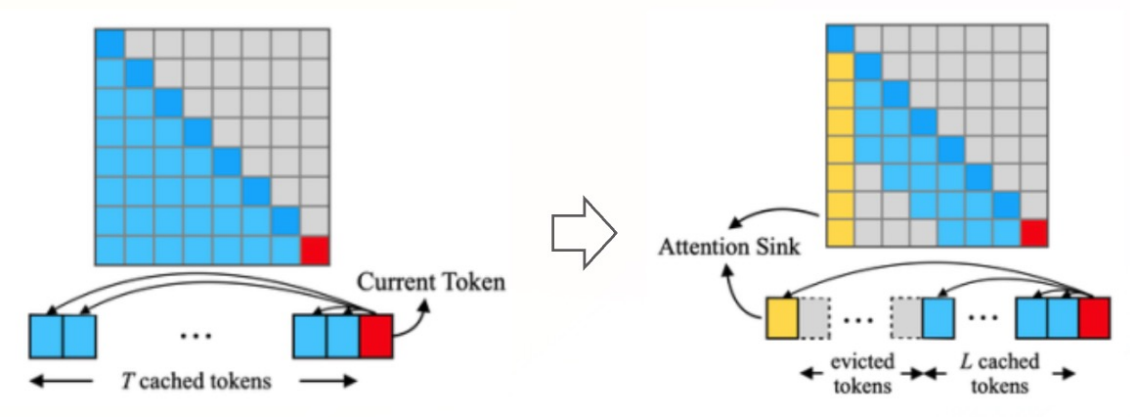


Benefits

- Significantly reduce the occupied activation video memory
- Enhance the performance of the available batch
- Quick Responses
- Alleviate the freezing issue during the decoding of extremely long texts

Solution2: Mamba

- Novel neural network
- Redefines sequence modeling by combining the strengths of state space models (SSMs)
- Mitigate limitations of Transformer models in handling long sequences efficiently



Solution3: Multi Block Decoding

- Parallelizes the decoding process across multiple "blocks" of tokens
- Breaks the sequential dependency of traditional autoregressive decoding while maintaining output coherence.
- Fully exploits SM parallel computing when seqlen reaches a certain threshold

HunYuan MoE Model Inference Acceleration – Scheduling Optimization

Ordinary Solution

➢ **Request Scheduling**

Only supports full scheduling or single step scheduling, with limited throughput/latency

➢ **Pull Scheduling**

Upstream cannot perceive the actual inference pressure of TensorRT-LLM workers on each node, and the inference load combined with the actual length of requests can easily cause service bottlenecks and time-consuming

Benefits

Benefits in Request

➢ Max batch size: 80 -> 115 (x**1.43**)

Benefits in Pull

➢ More balanced load and timely processing of requests

➢ More Flexible capacity expansion and contraction

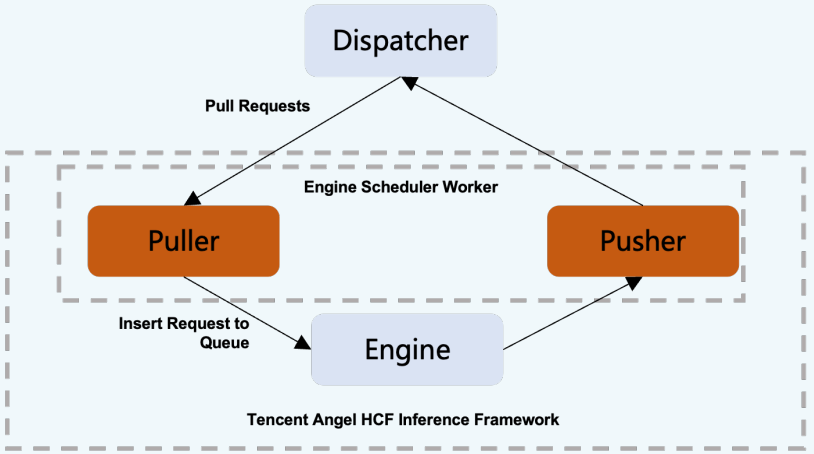
Our Solution

➢ **Request Scheduling**

Customize a performance first scheduling strategy based on the average output length as the scheduling decision

➢ **Pulling Scheduling**

- Change to pull mode
- TensorRT-LLM workers are stateless
- TensorRT-LLM workers actively pull requests from upstream based on their own inference load



HunYuan MoE Model Inference Acceleration –Weight Only Quantization

Weight Only Quant

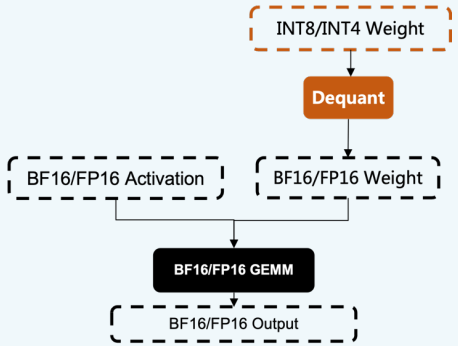
Weigh Only Quantization only reduces the numerical precision of to lower precision keeping activations and computations in higher precision.

Benefits

- Balances efficiency and accuracy
- Suitable for Memory Bound Model Infer

Challenges

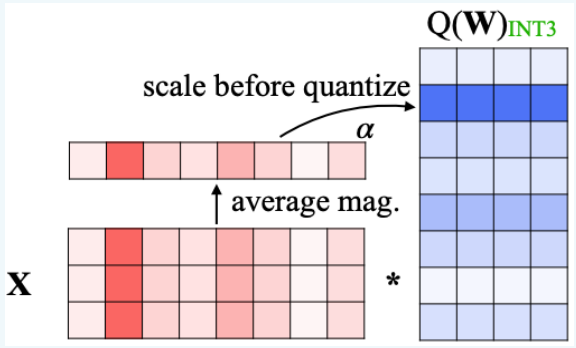
- Outlier Magnitude
- Channel Specific Sensitivity



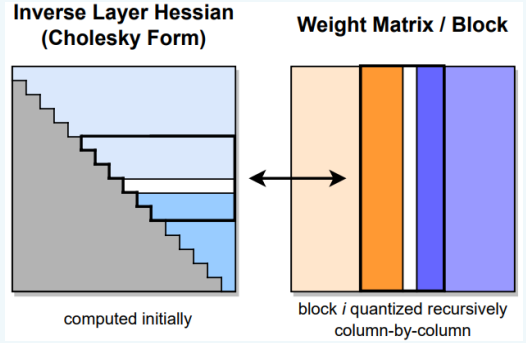
W8A16 | W4A16

Solutions

- Smooth Outliers by Using AWQ



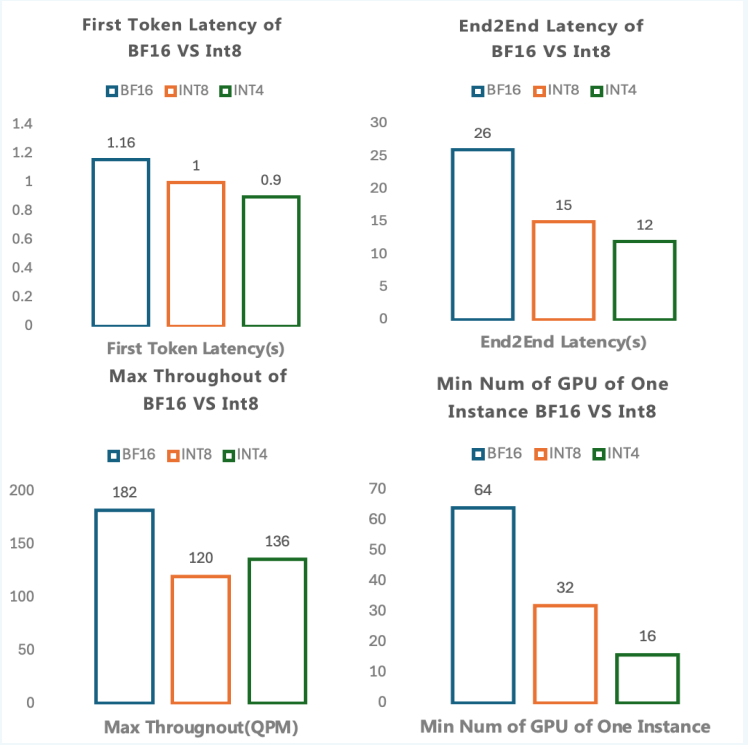
- Calibrate Weight Quantization Accuracy by Using GPTQ



- AWQ x GPTQ End-to-End Solution
  --Achieve Lossless Performance

Benefits

- **Benefits on Tencent Hunyuan Turbo**
  - First Token Latency: W8A16 reduced **16%**, W4A16 reduced **22%**
  - End2End Latency: W8A16 reduced **42%**, W4A16 reduced **54%**
  - Min Num GPU Required:W8A16 reduced **50%**, W4A16 reduced **75%**

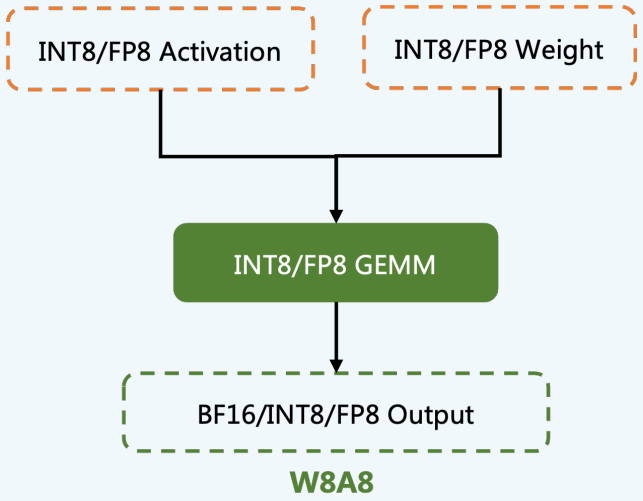


HunYuan MoE Model Inference Acceleration –Weight & Activation Quantization

Weight & Act Quant

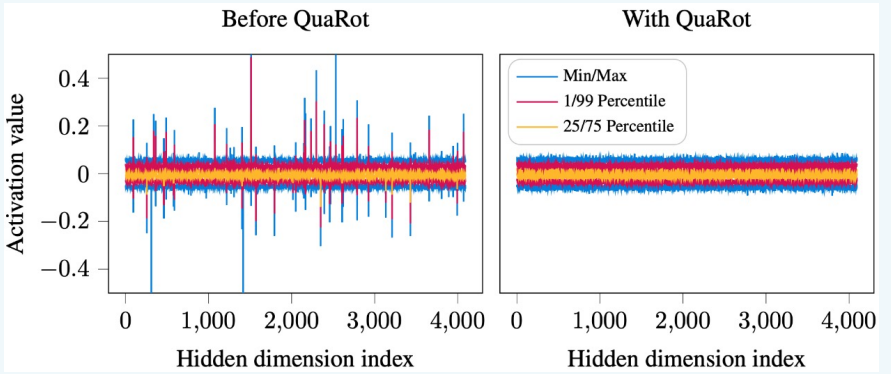
Weight and activation quantization reduces the numerical precision of both weight and activations.

- ➢Activation is hard to quant
- ➢Channel specific sensitivity



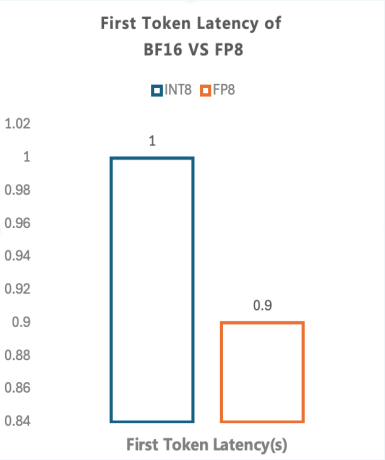
Solutions

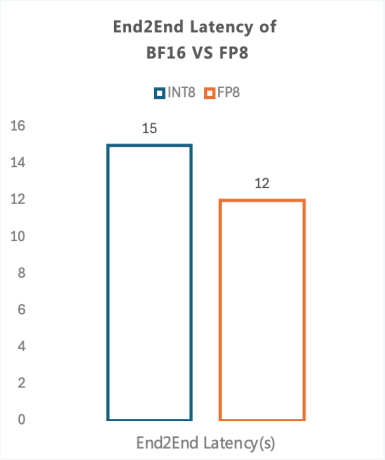
- ➢ Eliminate outliers activation using Hadamard Transform
- ➢ Obtain the Best Quantitative Scale with Grid Search and EMA
- ➢ Select high-quality PTQ calibration data by clustering sampling method



Benefits

- ➢ **Benefits on Tencent Hunyuan Dense, MoE, and Turbo-S**
  - First Token Latency: Reduced **10%**
  - End2End Latency: Reduced **20%**





HunYuan MoE Model Inference Acceleration –Lower Bit Quantization

Lower Bit Quant

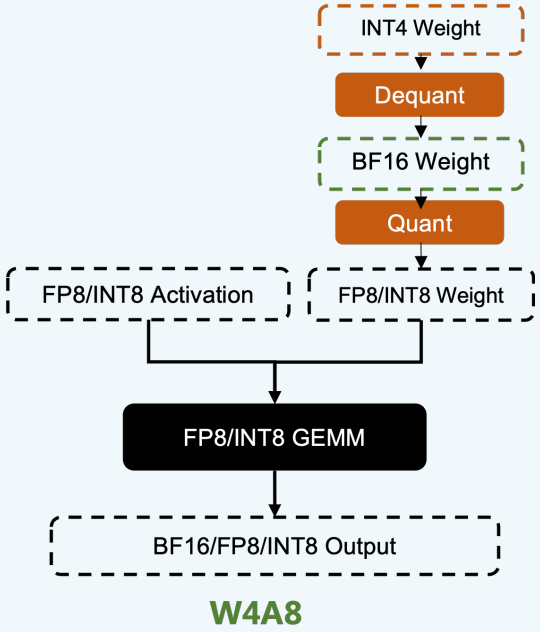
Weight and activation quantization reduces the numerical precision of both weight and activations to lower than 4 bit

.

Challenges

➢Outlier magnitude is extreme large on attention layer

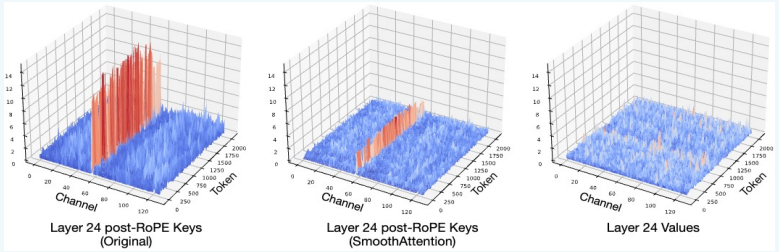
➢Channel specific sensitivity



W4A8

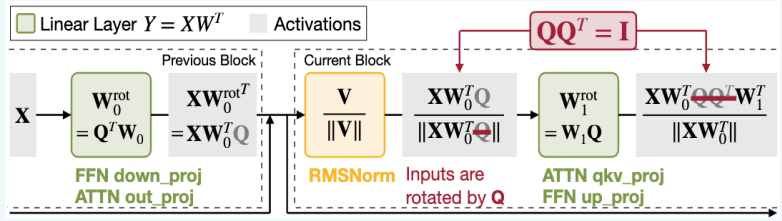
Solutions

➢ SmoothAttention: Eliminate outliers on the attention layer



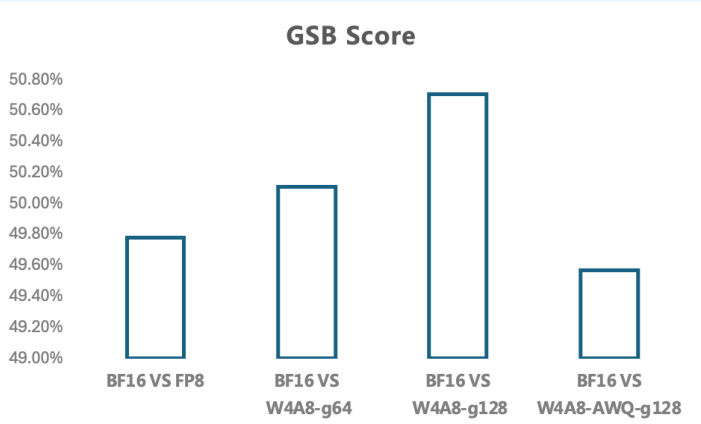
➢ Hadamard Matrix: Eliminate outliers in activation

➢ Compress weight into 4bit with AWQ



Benefits

➢ HunYuan trillion MoE results in lossless performance when gropu size=128 With AWQ



The higher the GSB winning rate, the better the former's performance

➢ HunYuan trillion MoE can be deploy on **single node**

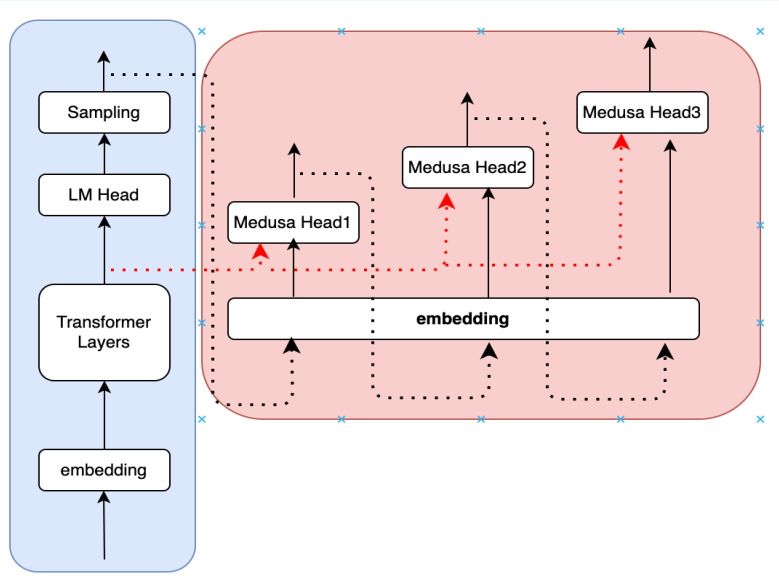
➢ First Token Latency reduces **15%**

➢ End2End Latency reduces over **10%**

HunYuan MoE Model Inference Acceleration –Speculative Sampling

Difficuties

Medusa Head are too simple, making it difficult to predict accurately after Head 2 on complex tasks



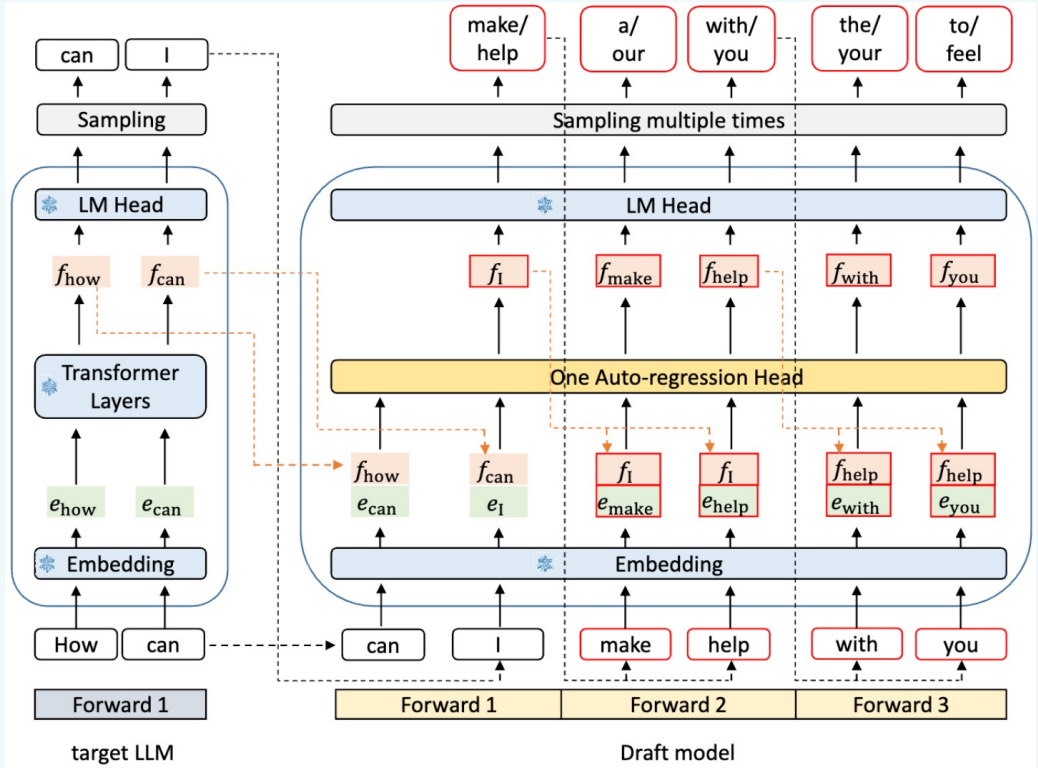
Benefits

Launch in Business :

- **Acceptance: 77%**
- **Generation Speed: 155 tokens/s -> 188 tokens/s** **17%**

Solutions

- Token-Level & Feature-Level Fusion Sampling
- Use Dynamic Draft Tree Generation Algorithm



Outline Overview

Introduction to Tencent's HunYuan Model and Taiji Angel

Optimization of inference in HunYuan large language models

Optimization of HunYuan Diffusion Model inference

Summary

HunYuan Diffusion Model Inference Acceleration

DiTs

Diffusion Transformers (DiTs)

- Transformer as the backbone network
- Enhanced Multi-Task Adaptability
- Architectural Scalability and Flexibility
- Scaling laws:Larger Model, Better Performance
- Skilled in Long-Range Dependencies within Image Tokens

Challenges

1. Large Model Parameters Leads to High Memory Usage

Take A Video Generation Model with 13B Param as Example:

- DiT Model Weight : 2x13x10^9bytes =24.2G
- Other Clip+MLLM+VAE models = 15G
- 24.2GB+15GB ≈ 40G

2. Long Seqence Length Requies Large Memory and Computations

- S=imgH/16*imgW/16*(frames/4+1)
- 720p-5s video, tokens=119056 ~ =120k
- Computational complexity Increase at the square level
- Activation of memory Increase at the square level

Single GPU torch	192x336x129f	720x1280x129f
50 step	1min30s	Estimated 1 hour
GPU Memory	45512M	OOM

Solutions

Algorithm x Engineering Acceleration

Video Memory Optimization

- Tensor Parallel TP → **Massive model parameters**
- \+ Sequential Parallel SP → **Long Sequence Problems**
- \+ Step distillation+CFG distillation → **Latency** -- Reduce Step Iterations and Batch Size
- \+ DeepCache acceleration → **Latency** -- Reduce Computation
- \+ FP8 quantization → **GPU Memory usage/Latency**
- \+ Sparse Attention → **Latency** -- Reduce Computation

Results

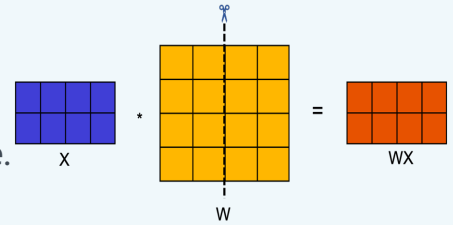
HunYuan Text-to-Video 720p	Torch flash single GPU	After optimizing the acceleration of multiple GPUs
Single inference time consumption	34.53s/iter	1.25s (x27.6)
End to end time consumption	1800s	37.5s (x48x)

HunYuan Diffusion Model Inference Acceleration-Tensor Parallel

Tensor Parallel

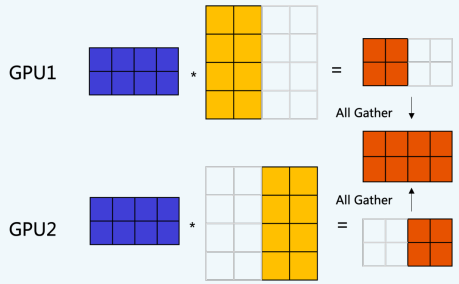
1. What’ s Tensor Parallel?

By dividing the weights and activations among devices, it enables faster training and inference.



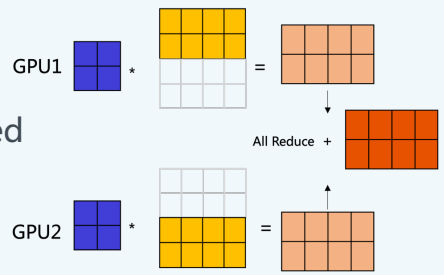
2. Column Parallel Linear Layer

- Splits the weight matrix of a linear layer by columns across multiple devices.
- The result needs to go through AllGather

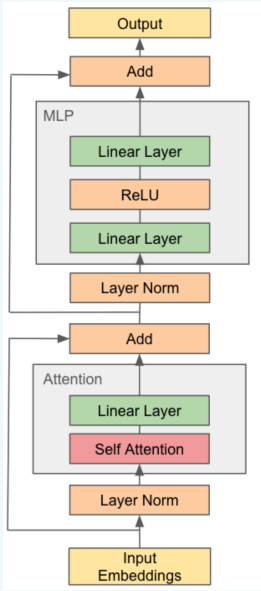


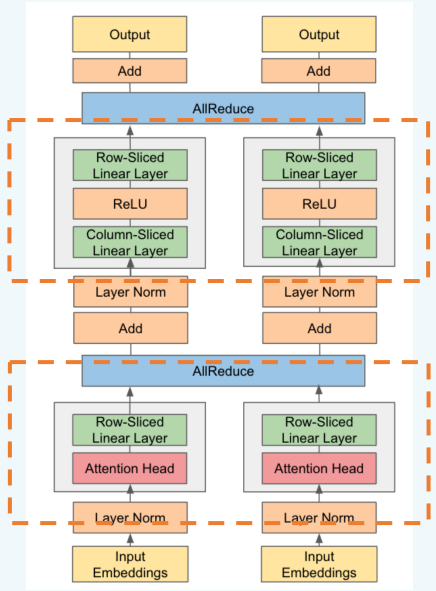
3. Row Parallel Linear Layer

- Each device processes the entire input data but only computes the output for its assigned rows of the weight matrix.
- The result needs to go through AllReduce



Our Tensor Parallel



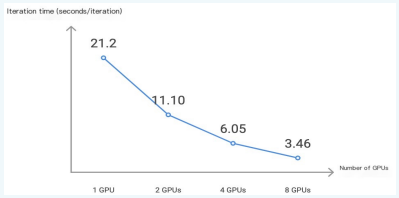


1. Combine MLP Tensor Parallel & AttentionTensor Parallel
2. Parallel on the Num of Head of Attention

- **Only Require Once AllReduce**
- **Minimize Communication Volume**

Results

2GPUs: **x1.9** 4GPUs: **x3.5** 8GPUs: **x6.18**



Limits

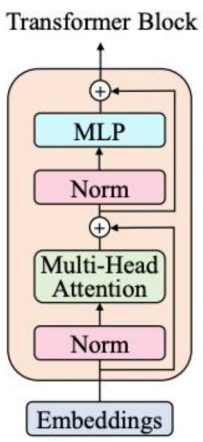
- Communication Overhead Increases with Number of GPU & Length of Seq
- Acceleration Gains is Nonlinear, Leading to Less Cost-Effectiveness.
- Leads to Memory Bottlenecks in long sequences scenerios

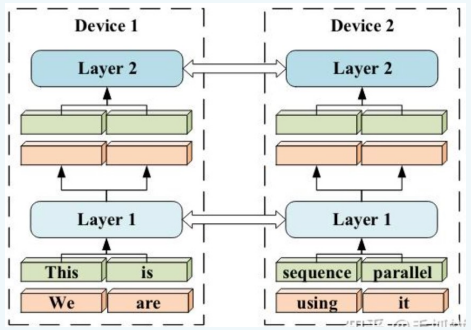
HunYuan Diffusion Model Inference Acceleration-Tensor Parallel x Seqence Parallel

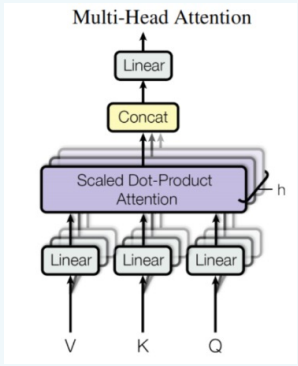
Challenges

What’s Seqence Parallel?

Splitting the sequence Length into multiple computing devices for parallel computing







Token Level Kernel

- No cross calculation is required between tokens, suitable for sequence parallelism.(eg. Linear，LayerNorm、MLP)

Attention Kernel

- Cross calculation is requied between tokens
- P-matrix after Softmax requires entire queries and values tensor
- Multiplication of the P-matrix by the value matrix also requires entire value tensor
- eg. MLA\MHA\GQA…

Seqence Parallel

DeepSpeed-Ulysses

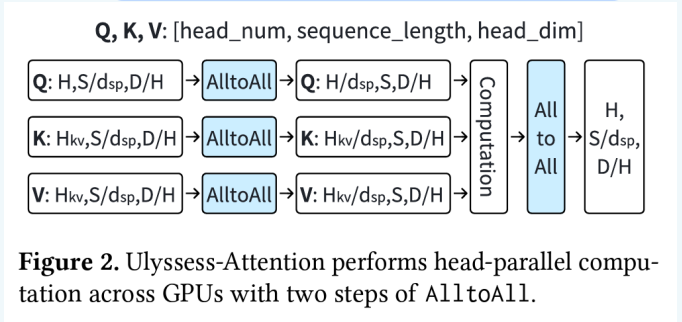
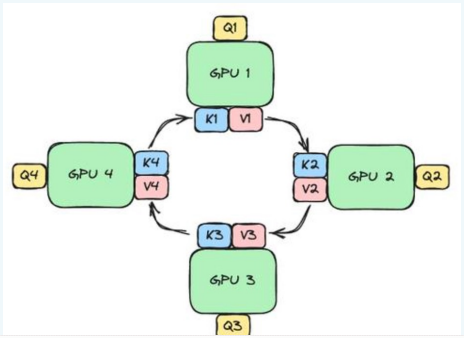
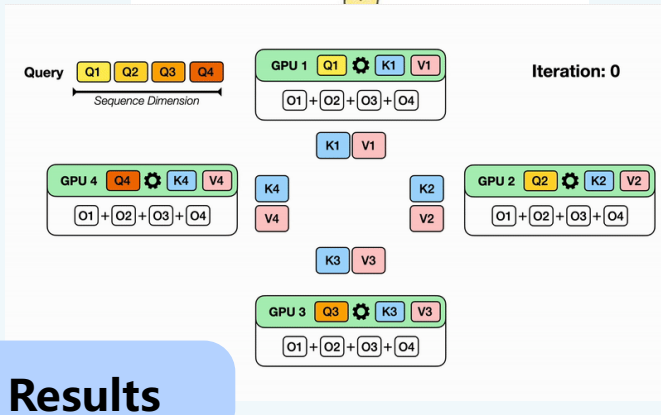


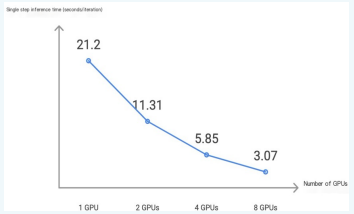
Figure 2. Ulyssess-Attention performs head-parallel computation across GPUs with two steps of `AlltoAll`.

Ring Attention





Results

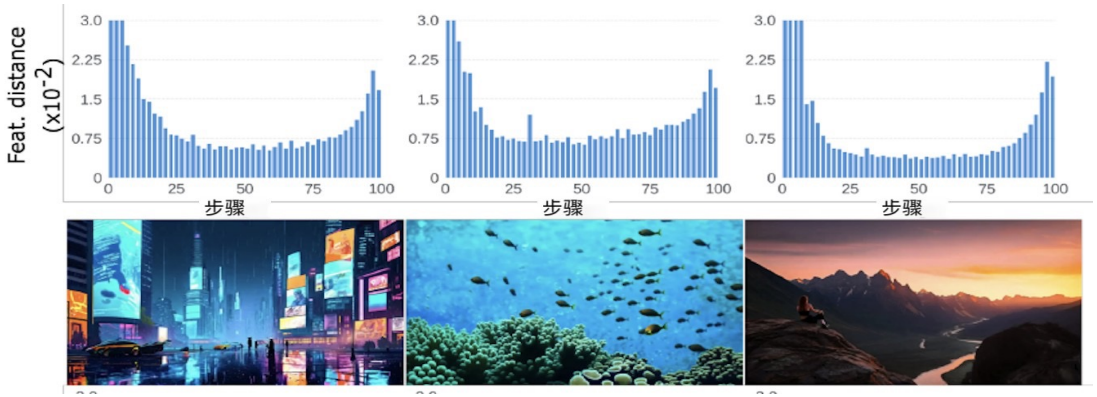


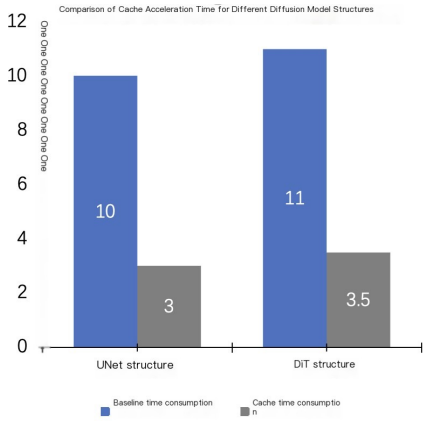
2GPUs: **x1.87** 4GPUs: **x3.62** 8GPUs: **x6.9**

As the number of GPUs increases, the acceleration yield increases with linear

HunYuan Diffusion Model Inference Acceleration-Activation Cache

Technical Approch





What's Activation Cache?

Utilize the high similarity of model features between adjacent steps in the diffusion model, cache some features, and reduce computational complexity

Benefits

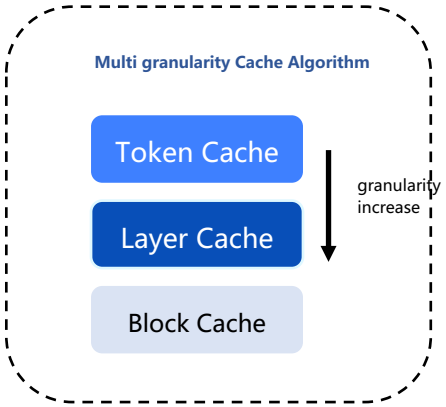
- Over **70+%** Acceleration without any Loss
- Training Free
- Excellent Adaptability, Suitable for T-I,T-V,T-3D

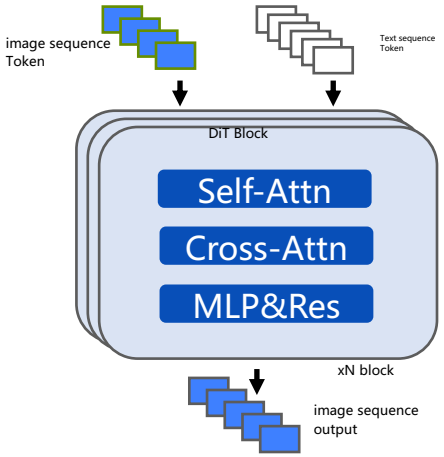
Solutions

Baseline

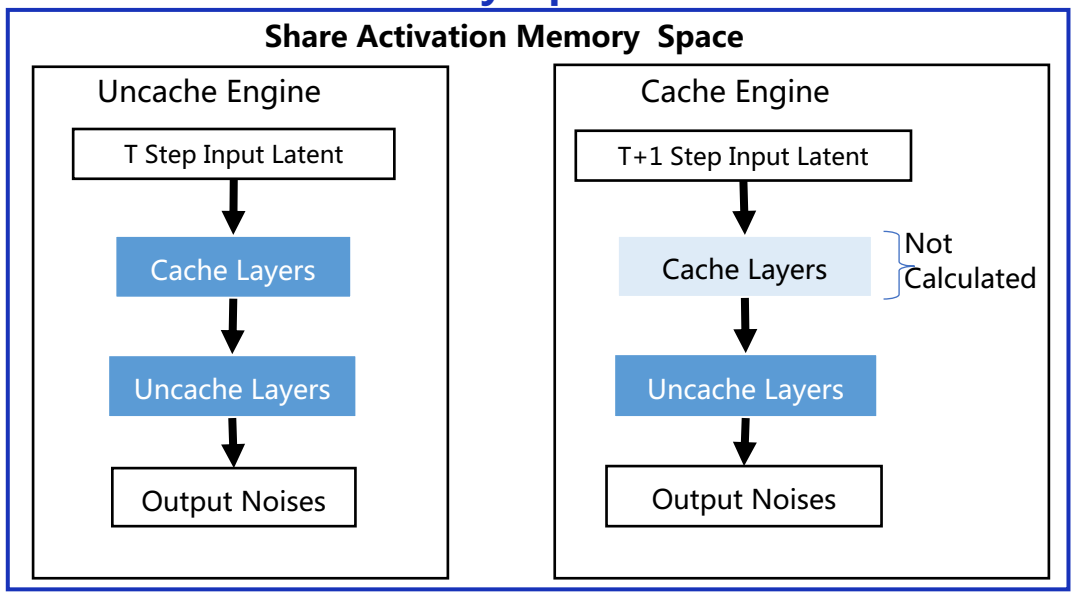
Cache

Algorithm optimization





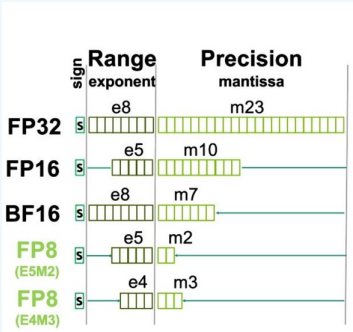
GPU Memory Optimization



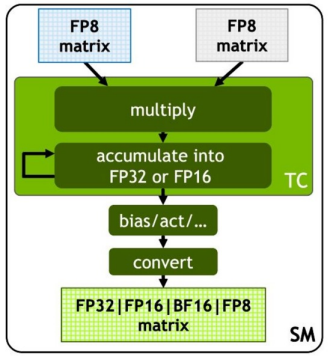
HunYuan Diffusion Model Inference Acceleration-FP8 Quantization

FP8

FP8 Quantization use 8-bit floating-point (FP8) numerical formats preserving the dynamic range and precision of floating-point arithmetic while using fewer bits, making it particularly effective for both training and inference.



Allocate 1 bit to either range or precision



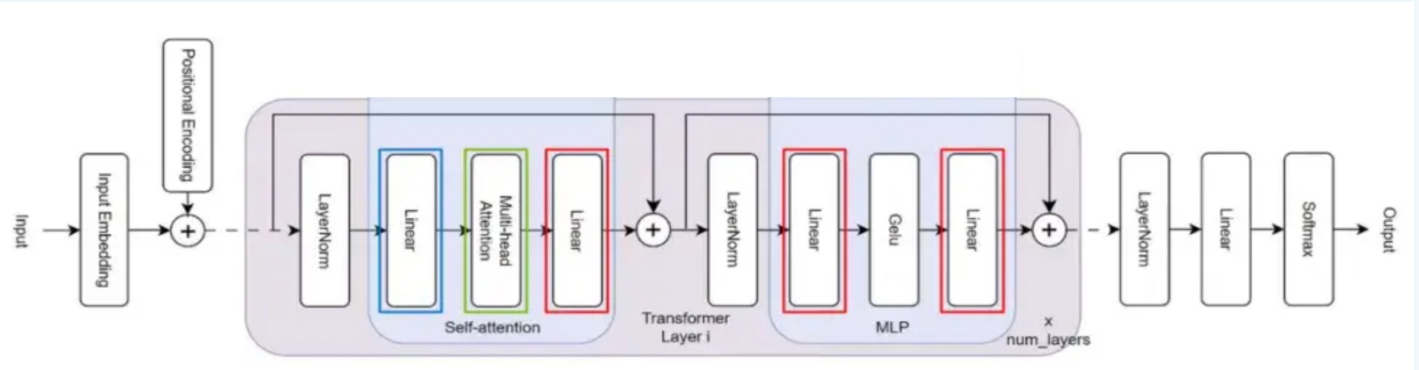
Support for multiple accumulator and output types

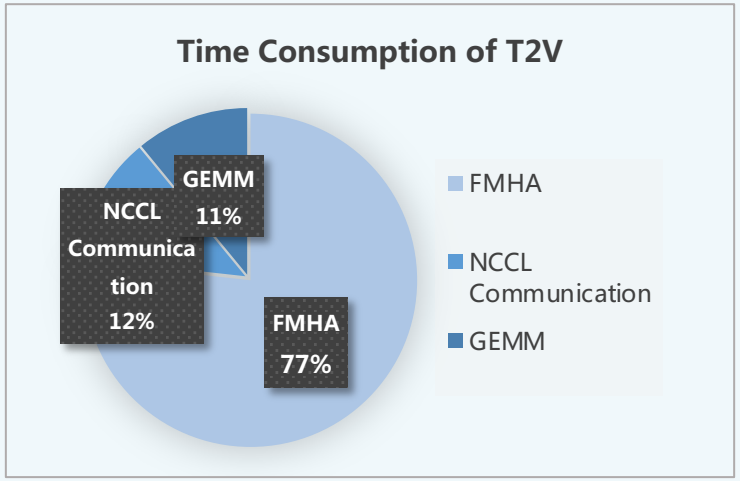
FP8 in T2V

- **QKV GEMM** : activation and weight
- **FMHA**: scaling factor=1
- **Other GEMM**: activation and weight

FMHA FP8 Quantification → **High** precision loss & **High** acceleration (**18%**)

GEMM FP8 Quantification → **Less** precision loss & **Less** acceleration (**5%**)



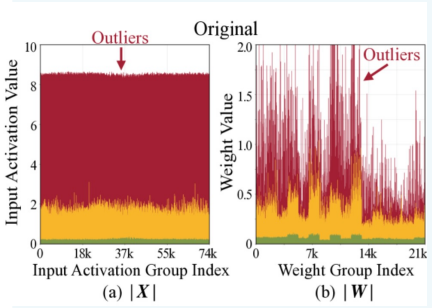


Bfloat16

GEMM fp8

HunYuan Diffusion Model Inference Acceleration-FP8 Quantization

Challenge



BF16 FP8

fp16

☒Hand Deform

Vanilla Quant

☒Blurriness

Almost Lossless

PTQ

Within Hours

Lossless

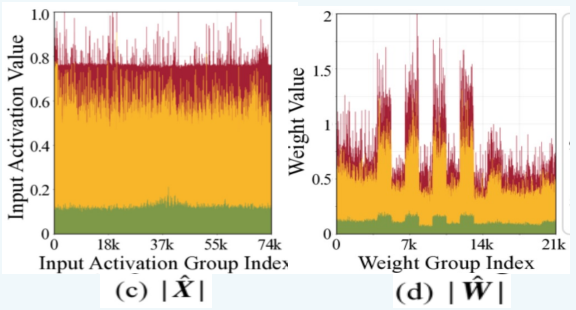
QAT

≈Cost of SFT

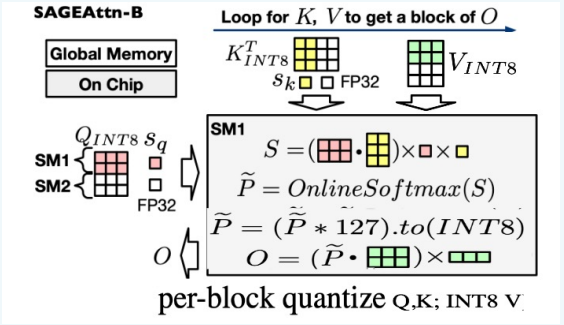
BF16 FP8

Technical Approch

1. Outlier/TimeStep smooth solves outliers



2. FMHA-per-Block quantization solves long seqlen quantization accuracy



Solution

- Sensity Layer/ Time Step Detection
- FMHA Per Block Quantization
- TimeStep Smooth
- Outlier Smooth

Results

✓ Hand

BF16 FP8

✓ Detailed

BF16 FP8

✓ Latency
✓ GPU Memory
✓ Lossless

Outline Overview

Introduction to Tencent's HunYuan Model and Taiji Angel

Optimization of inference in HunYuan large language models

Optimization of HunYuan Diffusion Model inference

Summary

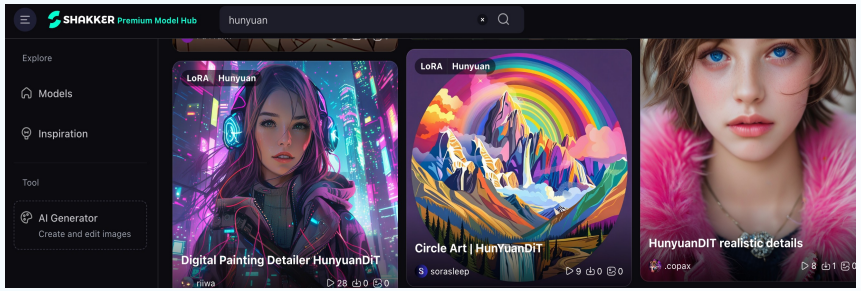
Tencent HunYuan AI Model Family Contribution in Open Source Community

HunYuan LLM

https://huggingface.co/tencent/Tencent-Hunyuan-Large
https://huggingface.co/tencent/Hunyuan-7B-Instruct

HunYuan Text-to-Image

https://huggingface.co/Tencent-Hunyuan/HunyuanDiT



HunYuan Text-to-Video

https://huggingface.co/tencent/HunyuanVideo

HunYuan Text-to-3D

https://huggingface.co/tencent/Hunyuan3D-2

NVIDIA Collaboration Summary and Prospects

Collaboration Summary

Adhere to the TensorRT-LLM Route

- From FT to TensorRT-LLM
- cuBLAS/cutlass
- Feed Back to the Open Source Community

Cooperation in Kernel Optimization

- Super Large Scale Mamba Kernel Optimization Solution
- Optimization for MoE FNN Communication

Multi-scene coverage

- LLM/Multimodal Understanding
- Text to Image/Video
- Text to 3D

Outlook for Cooperation

Participate in the co-construction of TensorRT-LLM

- Exploring usability and efficiency together
- HunYuan model version verification and automated verification mechanism
- Taiji AngelHCF corresponds to backend open source

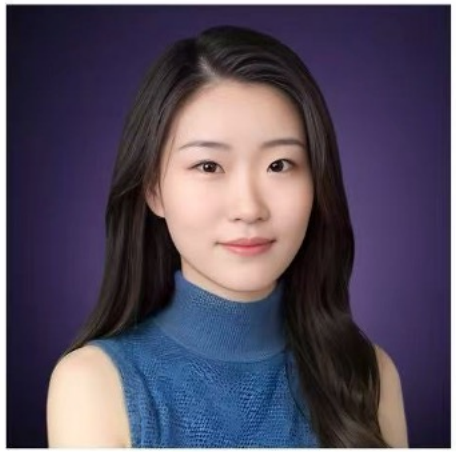
To Be Continued ...

NVIDIA GTC

3 月 20 日 06:00

AI 推理 / 推理微服务

[S71563] 基于 NVIDIA TensorRT-LLM 构建适用于大型模型的高性能推理引擎



孙艺芙

腾讯科技
高级应用研究员

王猛

NVIDIA
加速计算专家

Thank You !

Yifu Sun | Tencent

Meng Wang | Nvidia