

PoC 21: Validation of LLM for Network OAM Application on Generic Computing Platform


Rapporteur: China Telecommunications, Intel,
China Unicom, Huawei (UK) Technologies Ltd., CAICT

ENI PoC project #PoC 21: Validation of LLM for Network OAM Application on Generic Computing Platform



PoC Goals and PoC member task

Host/Team Leader:  - User Stories / Use Cases definition
- PoC development, documentation, and demos

Team members:  -Help with the architecture design, implementation of algorithm, testbed setup



-Participation in project discussions



- LLM system capability evaluation

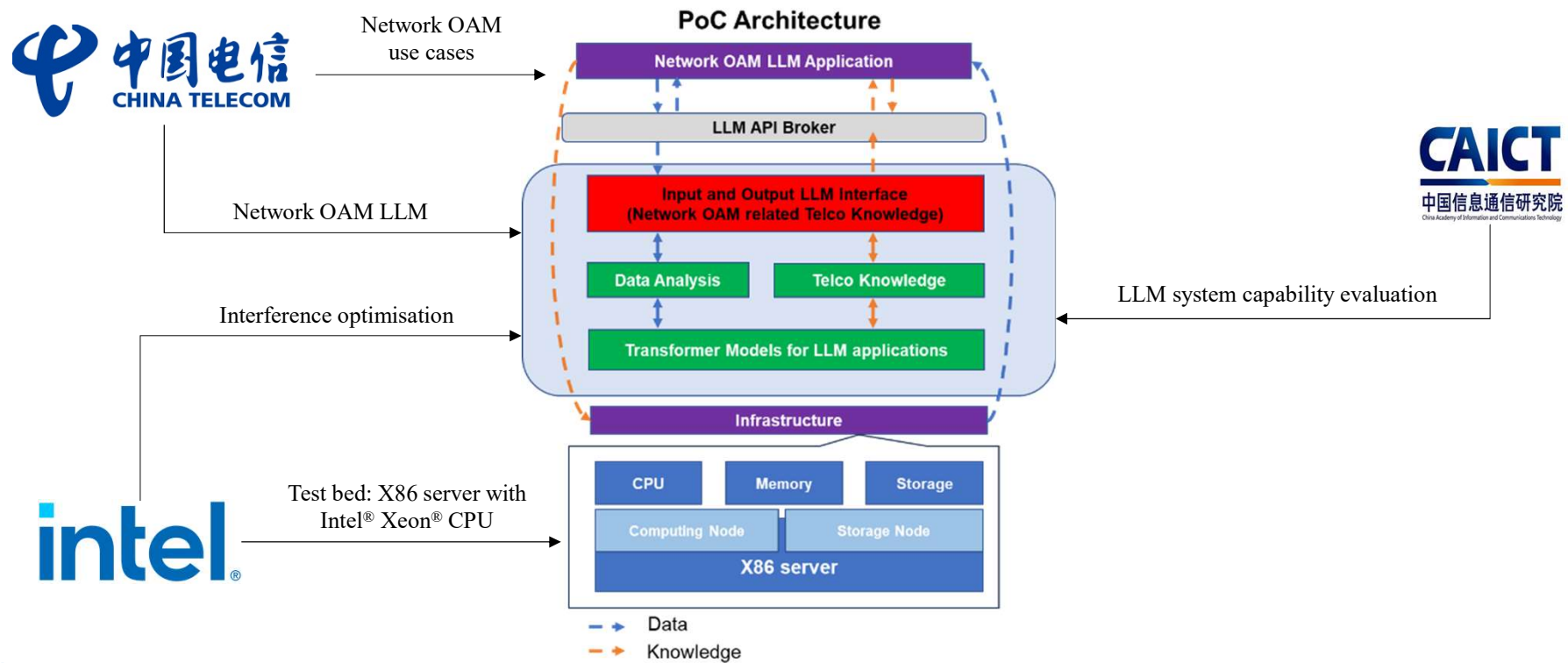
- ✓ • **PoC Project Goal #1:** AIGC application on X86 Platform. Demonstrate how to support AIGC application on X86 platform, support AIGC inference for LLM services, and realize the functionality and capability to provide identical AIGC request.
- ✓ • **PoC Project Goal #2:** Distributed AIGC computing optimization. Demonstrate organizing multiple computing node to provide a distributed solution for AIGC application.

ENI PoC project #PoC 21: Validation of LLM for Network OAM Application on Generic Computing Platform



PoC Architecture

The diagram represented below shows the framework of the PoC mapping to the ENI reference architecture.



ENI PoC project #PoC 21: Validation of LLM for Network OAM Application on Generic Computing Platform



Background: This PoC intends to demonstrate the feasibility and capability of network OAM LLM application running on generic computing platform aka X86 based platform instead of a GPU platform, with special attention to the lower cost and power consumption aspects, in the context defined by ENI.

In particular, this PoC solves the adaptation of generic computing platform e.g. X86 to replace part or all of a GPU platform for a reduced power and cost consumption. As demand for GPU increase dramatically due to the rise of AIGC applications, service providers having difficulties to provide sufficient computing power for AIGC related applications. The X86 platform on the other hand, cumulated large amount of computing power by stable service provider investment. The overall CPU workload for service provider often very low, less than 50%. To solve the unmatched problem of computing power and demand, it is a possible solution to employ AIGC applications on X86 servers. This PoC will demonstrate the computing ability by X86 platform, which realizes the balance between the CPU workload and lowering the power and budget cost.

ENI PoC project #PoC 21: Validation of LLM for Network OAM Application on Generic Computing Platform



Testbed Setup: X86 server with Intel® Xeon® CPU for LLM inference

CPU	5 th Gen Intel® Xeon® (Emerald Rapids)
Total Cores	52
Total Threads	104
Max Turbo Frequency	3.1Ghz
Processor Base Frequency	2.6 GHz
Cache	260MB

Test results with China Telecom's Network OAM LLM (13B)

I/O=1024/128			
Precision	First Token	Next Token	Throughput
BF16/FP16	540.75 ms	48.19 ms	20.75 tokens/s

```
question:请列举出中国电信相关的规章制度,包括但不限于但不限于:操作规程、应急预案、培训、设备维护保养、检查等。
<s> <human> :请列举出中国电信相关的规章制度,包括但不限于但不限于:操作规程、应急预案、培训、设备维护保养、检查等。
<bot> :以下是中国电信相关的规章制度:
1. 操作规程:如设备安装/退网、线路施工/维护、网络调整、故障处理等。
2. 应急预案:如自然灾害、网络事件、通信阻断、设备故障等。
3. 培训:包括职业安全、法律法规、产品技能、管理技能等方面的培训。
4. 设备维护保养:设备的维护规范、保养流程、故障处理等。
5. 检查:包括现场安全检查、网络运行质量检查、现场环境检查等。</s>
```


ENI PoC project #PoC 21: Validation of LLM for Network OAM Application on Generic Computing Platform



PoC Milestones and Current Progress



PoC Milestone	Stages/Milestone description	Target Date	Additional Info
P.S	PoC project submission	03/2024	Proposal approved for PoC reviewing during #ENI 29
P.D1	PoC Demo 1	06/2024	Demo in ENI #30
P.D2	PoC Demo 2	09/2024	Demo at ENI#31
P.C1	PoC Expected Contribution 1	10/2024	contributions to ENI requirements.
P.C2	PoC Expected Contribution 2	10/2024	contributions to ENI use case.
P.R	PoC Report	12/2024	PoC-Project-End Feedback
P.E	PoC Project End	01/2025	Presented to ISG ENI for information