Demonstrating a Deployable Optical Inference Server for Billion-Parameter-Scale LLMs

Lumai: Phillip Burr, Xianxin Guo, James Spall. Oxford, UK Contact: phillip.burr@lumai.ai

Key Result: A fully integrated optical AI server performing end-to-end inference of billion-parameter LLMs. Suitable for low-cost, high-throughput inference applications. Compatibility with standard data center infrastructure.

THE INFERENCE ERA SCALING PROBLEM

Foundational AI target: ~1000× increase in compute over the next five years. However scaling with conventional digital accelerators would imply ~ $100T investment and ~1000 GW of power. AI inference is constrained by power.

A FUNDAMENTALLY DIFFERENT APPROACH

Incremental improvements in silicon come with a significant power cost. Architectural specialization (GPUs, tensor processors, low precision formats) is approaching its limits. **Sustained scaling of inference requires new technology approaches**.. Optical compute provides a path to scaling AI in the inference era.

OPTICAL MATRIX MULTIPLICATION

Input vectors are encoded in light, replicated via optical propagation. Weights are applied using programmable optical elements. Outputs are summed optically in a single step.

Key property:

- Compute scales with spatial parallelism
- Operations scale $\propto N^2$
- Energy scales $\propto N$

LUMAI IRIS SYSTEM ARCHITECTURE

A complete AI inference server integrating:

- Digital subsystem - control, memory, networking. Compatible with standard AI frameworks
- Optoelectronic Interface - high-speed conversion
  Optical Tensor Engine - executes matrix multiplication in optical domain

Each matrix multiplication is completed in a single cycle

Deployment

- Standard rack integration
- Air-cooled data center compatibility

SYSTEM VALIDATION - LUMAI IRIS NOVA:

- First full-server deployment of optical compute for billion parameter LLM inference
- Scalable architecture targeting 2048 × 2048+

System Integration

- Stable operation within standard data center environments

End-to-End Inference

- Billion-parameter LLM inference demonstrated (Llama)
- Real-world workload execution

CONCLUSION

- Free-space optical computing can be integrated into a standard server architecture
- Demonstrated billion-parameter LLM inference
- Provides a practical path to scaling AI inference beyond digital hardware limits
- Optical AI acceleration enables substantial improvements in energy efficiency while maintaining deployability in real data center environments

LUMAI

lumai.ai