AI Software and the GPU Industry

A Symbiotic Evolution

June 2025 · Dinesh

The GPU-AI Foundation

Why GPUs Power AI:

- Thousands of CUDA cores vs. CPU's ~10-64 cores
- Massive parallelism for tensor operations
- Optimized for matrix multiplies and convolutions
- High memory bandwidth (HBM3, GDDR6X)

Software Dependencies:

- Frameworks rely on GPU libraries (cuDNN, cuBLAS)
- Tensor Cores accelerate mixed-precision compute
- SIMT execution model matches AI workloads

Software Drives Hardware Innovation

AI Software Explosion:

- Large models: GPT, Gemini, Claude demand massive compute
- Frameworks push GPU limits: DeepSpeed, vLLM, Triton
- Edge AI opens new markets: Jetson, CoreML

Software Stack Impact:

Layer	Examples	GPU Impact
Frameworks	PyTorch, JAX	Dynamic compute graphs
Compilers	XLA, Triton	Kernel fusion optimization
Inference	TensorRT, vLLM	Latency-optimized compute
Infrastructure	Ray, KServe	Multi-GPU scalability

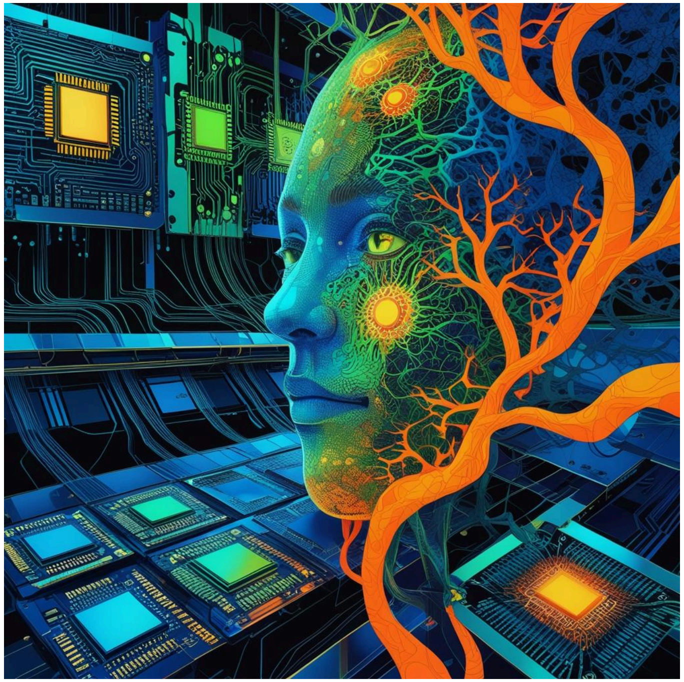
Hardware-Software Co-Evolution

Feedback Loop:

```
AI Software Needs → GPU Architecture Changes
     ↑                        ↓
Performance Bottlenecks ← New Hardware Features
```

Real Examples:

- **Mixed-precision:** FP16, bfloat16, FP8 support
- **Communication:** NCCL, NVLink for multi-GPU
- **Memory:** SRAM improvements for transformer models
- **Specialized units:**



Future Trends & Summary

Emerging Trends:

- Open-source GPU stacks (ROCm, Triton)
- Multi-backend compilers (IREE, TVM)
- Cross-hardware abstractions
- Energy-efficient "Green AI"

Key Takeaways:

- AI software and GPUs are deeply interdependent
- Software innovation drives GPU adoption and design
- GPUs enable software breakthroughs through scale
- Co-evolution defines the performance frontier

Thank You

Dinesh

✉ dineshkumarb@gmail.com

🌐 https://dkbhaskaran.github.io/

📞 +1 999 999 9999