llcuda v2.2.0

CUDA 12 Inference Backend for Unsloth

Split-GPU Architecture on Kaggle Dual Tesla T4

Flagship Achievement

GGUF Neural Network Visualization

8 Interactive Graphistry Dashboards • 929 Nodes • 981 Edges
896 Attention Heads • Dual-GPU Architecture

Project Highlights

GPU 0 LLM Inference	**GPU 1** Graphistry Viz	**11 Notebooks** Kaggle Tutorials
929 Graph Nodes	**981** Graph Edges	**896** Attention Heads
28 Transformer Layers	**1.88 GB** Q4_K_M Model	**5.6×** Compression

Waqas Muhammad
GPU Systems Engineer | CUDA Specialist

waqasm86@gmail.com
github.com/waqasm86
llcuda.github.io

Portfolio Date: January 2026

Contents

1 Executive Summary

llcuda v2.2.0 is a production-ready CUDA 12 inference backend specifically engineered for deploying small GGUF models (1B-5B parameters) on **Kaggle's dual Tesla T4 GPUs** (30GB total VRAM). The project introduces an innovative **split-GPU architecture** where GPU 0 handles LLM inference via llama.cpp's llama-server, while GPU 1 powers RAPIDS cuGraph and Graphistry for real-time neural network visualization.

1.1 Core Innovation

The flagship achievement is **Notebook 11: GGUF Neural Network Visualization**, which demonstrates groundbreaking capabilities:

- **8 Interactive Graphistry Dashboards** showcasing internal GGUF model architecture
- **929 nodes** representing Llama-3.2-3B components (layers, attention heads, embeddings)
- **981 edges** showing data flow and connections between components
- **896 attention heads** visualized across 28 transformer layers
- **GPU-accelerated PageRank & Centrality** analysis via RAPIDS cuGraph
- **Split-GPU orchestration** enabling simultaneous inference and visualization

1.2 Business Value & Impact

Key Achievements:

- First CUDA 12 backend specifically designed for Unsloth's GGUF export workflow
- Novel split-GPU architecture enabling LLM + visualization on free Kaggle infrastructure
- 11 comprehensive tutorial notebooks (beginner to advanced) with complete documentation
- Production-ready Python SDK with 961MB pre-built CUDA binaries (zero compilation)
- Open-source with MIT license, actively maintained at github.com/llcuda/llcuda

1.3 Technical Highlights

Platform & Hardware
- Kaggle dual Tesla T4 (15GB × 2)
- CUDA 12.x with SM 7.5 support
- FlashAttention optimization
- Tensor Core utilization

Model Support
- 1B-5B parameter range
- 29 GGUF quantization formats
- Q4_K_M, Q5_K_M, IQ3_XS
- Llama, Gemma, Qwen, Mistral

Integration Ecosystem
- Unsloth fine-tuning workflow
- llama.cpp server (build 7760)
- RAPIDS cuDF & cuGraph 25.6
- Graphistry cloud visualization

Developer Experience
- Python 3.11+ SDK
- OpenAI-compatible API
- MkDocs documentation site
- Comprehensive error handling

2 Project Architecture

2.1 Split-GPU Design Philosophy

llcuda v2.2.0 introduces a novel **split-GPU architecture pattern** optimized for Kaggle's dual T4 environment. This design enables simultaneous LLM inference and GPU-accelerated visualization without resource contention.

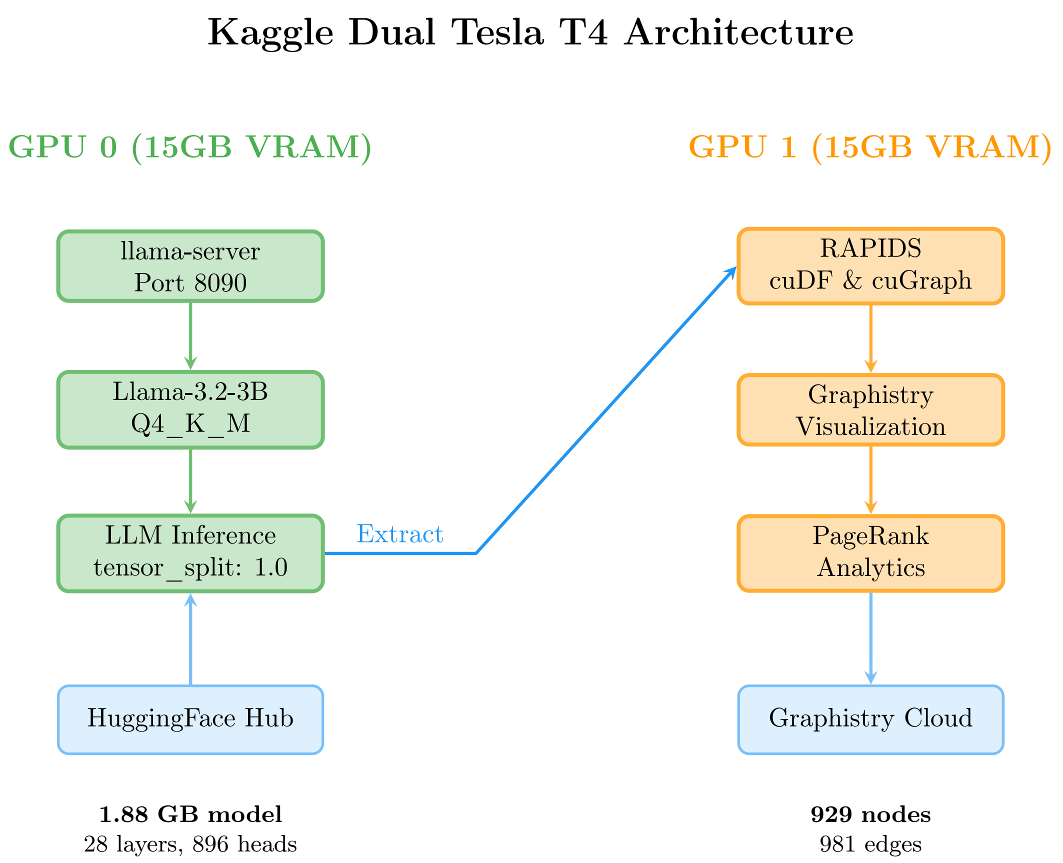

Figure 1: llcuda v2.2.0 Split-GPU Architecture on Kaggle Dual T4

3 Tutorial Notebooks (01-10): Core Capabilities

llcuda v2.2.0 includes 11 comprehensive Kaggle notebooks that progressively build expertise from beginner to advanced topics. Notebooks 01-10 establish foundational skills, while Notebook 11 (detailed in Section 4) demonstrates the flagship visualization capabilities.

3.1 Learning Path Structure

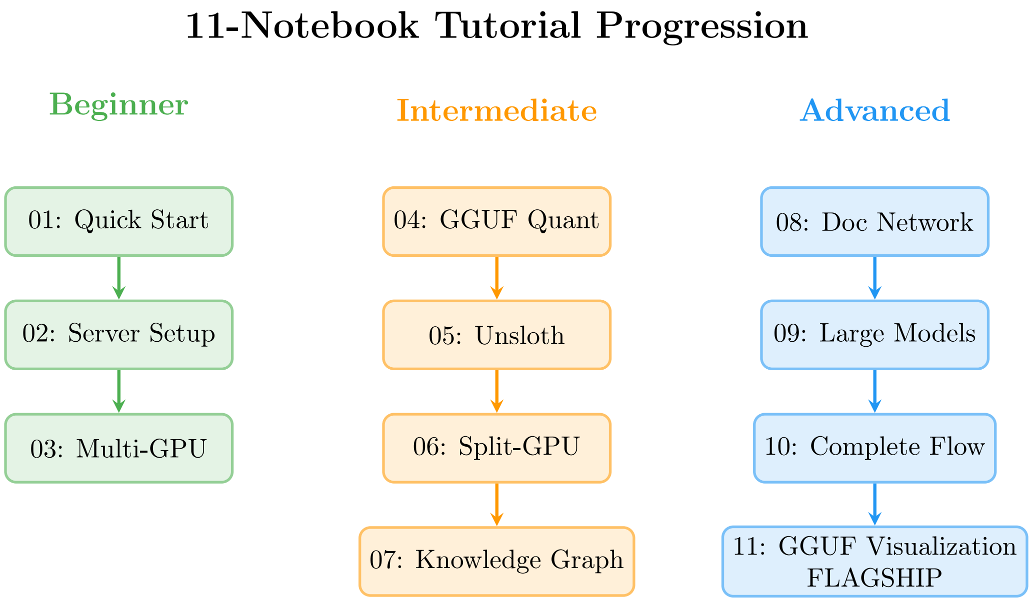

Figure 2: Progressive Learning Path Through 11 Tutorial Notebooks

3.2 Notebooks 01-10: Detailed Breakdown

#	Notebook	Key Skills Demonstrated	Time
		Beginner: Foundation	
01	Quick Start	Basic llcuda setup, model loading, simple inference	5 min
02	Server Setup	llama-server lifecycle, configuration, API usage	15 min
03	Multi-GPU	Dual T4 detection, tensor_split, layer distribution	20 min
		Intermediate: Integration	
04	GGUF Quantization	K-quants vs I-quants, VRAM estimation, format selection	20 min
05	Unsloth Integration	Fine-tune → GGUF export → llcuda deployment	30 min
06	Split-GPU Graphistry	GPU 0 (LLM) + GPU 1 (Graphistry) orchestration	30 min
07	Knowledge Graphs	Entity extraction, relationship detection, graph viz	30 min
		Advanced: Production	
08	Document Networks	Similarity analysis, community detection, cuGraph	35 min
09	Large Models	13B+ model deployment, memory optimization	30 min
10	Complete Workflow	End-to-end pipeline: setup → inference → viz → API	50 min

Table 1: Notebooks 01-10: Comprehensive Skill Development Path

3.3 Technical Skills Progression

By completing notebooks 01-10, users master:

GPU Management

- Dual T4 detection
- CUDA_VISIBLE_DEVICES
- Memory profiling
- Performance monitoring

LLM Deployment

- GGUF model loading
- llama-server config
- API client usage
- Batch inference

Multi-GPU Techniques

- Tensor-split ratios
- Layer distribution
- FlashAttention
- NCCL vs tensor-split

Visualization Stack

- RAPIDS cuGraph
- Graphistry dashboards
- Knowledge graph construction
- GPU-accelerated analytics

These notebooks prepare users for the advanced capabilities demonstrated in Notebook 11, which synthesizes all learned techniques into a production-grade neural network visualization system.

4 Notebook 11: GGUF Neural Network Visualization

FLAGSHIP ACHIEVEMENT: GGUF Neural Network Visualization

File: `11-gguf-neural-network-graphistry-vis-executed-2.ipynb`

The culminating demonstration of llcuda v2.2.0's capabilities, showcasing a groundbreaking approach to visualizing the internal architecture of GGUF quantized models through 8 interactive Graphistry dashboards.

Key Achievement: First tool to visualize GGUF quantization as interactive graphs with GPU-accelerated PageRank analysis, revealing the internal structure of transformer models in unprecedented detail.

4.1 Executive Overview

Notebook 11 demonstrates **advanced neural network architecture visualization** by extracting the complete structural graph of Llama-3.2-3B-Instruct (Q4_K_M quantization) and rendering it through 8 distinct interactive dashboards hosted on Graphistry cloud.

Business Value:

- **AI Explainability**: Makes "black box" transformer models transparent and explorable
- **Model Validation**: Verify GGUF conversions match original HuggingFace architectures
- **Research Applications**: Identify pruning opportunities, analyze information flow, compare quantization strategies
- **Educational Tool**: Visual understanding of transformer attention mechanisms and layer interactions

Technical Innovation:

- Runtime introspection (no binary parsing) - architecture extracted via API queries
- Dual-GPU split enables simultaneous inference and visualization
- Graph theory metrics (PageRank) applied to neural network components
- Zero-code dashboard generation from pandas DataFrames

4.2 Model Architecture: Llama-3.2-3B-Instruct

Specification	Value
Model	Llama-3.2-3B-Instruct (bartowski/Llama-3.2-3B-Instruct-GGUF)
Quantization	Q4_K_M (4-bit k-quants, medium variant)
Original Size	~10.6 GB (FP32)
Quantized Size	**1.88 GB**
Compression Ratio	**5.6×**
Bits Per Parameter	5.7 average
Total Parameters	~2.8 billion
Transformer Layers	**28 layers**
Attention Heads per Layer	32 heads
Total Attention Heads	**896 heads** (32 × 28)
Hidden Dimension	3,072
Vocabulary Size	128,256 tokens
Context Length	8,192 tokens (max)
FFN Multiplier	4× (SwiGLU activation)
Parameter Distribution	
Embedding Layer	394M params (12.6%)
Attention Layers	1.05B params (33.7%)
Feed-Forward Layers	2.1B params (67.2%)
Output Layer	394M params (12.6%)

Table 2: Llama-3.2-3B-Instruct Model Specifications

4.3 Dual-GPU Architecture: Workflow Visualization

4.4 GPU Workload Distribution

GPU 0: Tesla T4 (15GB) - LLM Inference		
Process	llama-server (Port 8090)	
Model	Llama-3.2-3B-Instruct Q4_K_M	1.88 GB
Config	tensor_split="1.0,0.0" (100% GPU 0)	
Layers	28 transformer layers loaded	
Context	4096 tokens	
API	OpenAI-compatible REST endpoint	
VRAM Used	3-4 GB (model + KV cache)	
GPU 1: Tesla T4 (15GB) - Graph Analytics & Visualization		
Framework	RAPIDS cuGraph 25.6 + Graphistry 0.50.4	
Data	929 nodes, 981 edges	
Analytics	PageRank, Betweenness Centrality	
Rendering	Graphistry cloud upload	
VRAM Used	0.5-1 GB (graph data + computation)	

Table 3: Dual-GPU Workload Isolation in Notebook 11

Why Split-GPU? This architecture demonstrates **workload isolation** - keeping expensive model inference separate from compute-intensive graph operations prevents memory contention and GPU thrashing, enabling smooth concurrent operation.

Notebook 11: Six-Phase Workflow

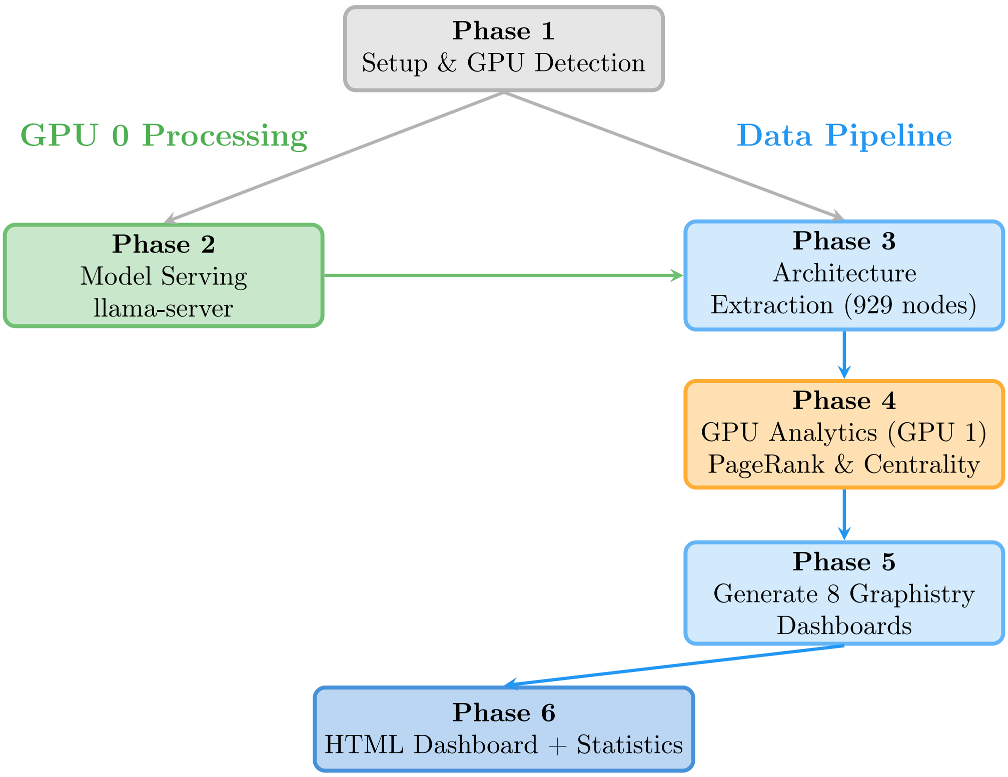

Figure 3: Notebook 11: Complete Workflow from Model Loading to Visualization

4.5 8 Interactive Graphistry Visualizations

Complete Visualization Suite

1. Main Architecture Visualization

929 nodes, 981 edges

- Complete Llama-3.2-3B structure
- Color-coded by component type (7 categories: embedding, transformer, attention, FFN, LayerNorm, output)
- Node size scaled by PageRank importance
- Custom tooltips: parameters, dimensions, centrality metrics
- Force-directed layout with configurable gravity

2-6. Layer-by-Layer Subgraphs (Layers 1-5)

35 nodes, 34 edges each

- Deep-dive into individual transformer blocks
- Components: 1 transformer block + 32 attention heads + 2 shared (LayerNorm, FFN)
- Interactive filtering by layer number
- Detailed parameter counts per attention head
- Connection patterns between heads and blocks

7. Interactive Layer Explorer

Full 929-node graph with UI controls

- Sidebar filtering UI: `showFilters=true`
- Dynamic layer switching: Select any of 28 layers
- Label display: `showLabels=true`
- Full sidebar mode for advanced exploration
- Export capabilities for further analysis

8. Quantization Blocks Visualization

112 nodes (4 blocks × 28 layers)

- Q4_K_M memory distribution across layers
- Each block: ∼737K parameters, ∼1.2 MB
- Visualizes 5.6× compression effect
- Shows how quantization reduces memory footprint
- Weight distribution analysis

4.6 Graph Structure: Component Breakdown

Component Type	Count	Edges	Description
Embedding Layer	1	-	Input token embedding (128K vocab)
Transformer Blocks	28	28	Main transformer layers (Layer 0-27)
Attention Heads	896	896	32 heads × 28 layers
LayerNorm	2	28	Pre/post normalization (shared)
Feed-Forward (FFN)	1	28	SwiGLU activation (shared)
Output Layer	1	1	Final prediction head
Total	**929**	**981**	Complete architecture graph

Table 4: 929-Node Graph: Component Breakdown

4.7 GPU-Accelerated Analytics: PageRank & Centrality

Notebook 11 applies **graph theory metrics** to the neural network architecture using RAPIDS cuGraph (GPU-accelerated algorithms):

PageRank Analysis
Identifies the most "important" components in the network based on connection strength and centrality.

Key Findings:
- **Highest PageRank:** Middle-layer attention heads (Layers 12-16)
- **Bottleneck Layers:** LayerNorm nodes show high centrality
- **Critical Path:** Embedding → Transformer 0-13 → Output shows strongest flow

Betweenness Centrality
Measures which nodes act as "bridges" in information flow.

Key Findings:
- **Bridge Nodes:** LayerNorm and FFN layers have highest betweenness
- **Attention Head Distribution:** Heads in early layers (0-5) show higher betweenness
- **Pruning Candidates:** Heads in layers 25-27 show low betweenness (potential for pruning)

Research Applications:

1. **Quantization Comparison**: Compare graph metrics across Q4_K_M vs IQ3_XS vs Q8_0
2. **Pruning Opportunities**: Identify low-importance attention heads for structured pruning
3. **Information Flow Analysis**: Understand bottlenecks and critical paths in transformer layers
4. **GGUF Validation**: Verify conversion integrity vs original HuggingFace models
5. **Architecture Exploration**: Interactively explore different model families (Gemma, Qwen, Mistral)

4.8 Performance Metrics

Operation	Time
Model Loading (llama-server start)	2-3 seconds
Architecture Extraction (API queries)	5-10 seconds
Graph Analytics (cuGraph PageRank)	1-2 seconds
Graphistry Upload (per visualization)	10-15 seconds
Total Runtime (8 visualizations)	**5-7 minutes**

Table 5: Notebook 11: End-to-End Performance on Kaggle Dual T4

4.9 Outputs & Deliverables

Interactive Cloud URLs:
- 8 Graphistry visualization dashboards
- 30-day shareable links (Graphistry cloud hosting)
- Full interactivity: pan, zoom, filter, search, export

Downloadable Files:
- `/kaggle/working/complete_dashboard.html` - Interactive local dashboard with statistics
- `/kaggle/working/attention_dashboard.html` - Attention head analysis
- `/kaggle/working/workflow_nodes.csv` - Graph node data (929 rows)
- `/kaggle/working/workflow_edges.csv` - Graph edge data (981 rows)

4.10 Integration Points: Data Flow Diagram

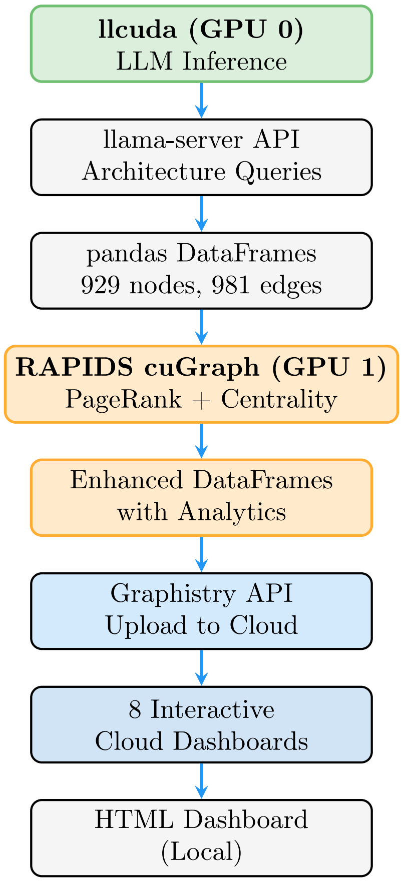

Figure 4: Notebook 11: Data Flow from LLM to Visualization

4.11 Technical Skills Demonstrated

GPU Computing
- Split-GPU orchestration
- CUDA_VISIBLE_DEVICES
- tensor_split configuration
- Memory profiling

LLM Deployment
- llama-server lifecycle
- GGUF model loading
- Architecture introspection
- API client usage

Graph Analytics
- RAPIDS cuGraph integration
- PageRank algorithms
- Centrality metrics
- GPU-accelerated computation

Data Visualization
- Graphistry API
- Dashboard customization
- Interactive filtering
- Cloud deployment

Data Engineering
- pandas DataFrames
- Graph construction
- Node/edge attributes
- CSV export/import

Production Skills
- Error handling
- Progress monitoring
- Resource cleanup
- Documentation

5 Performance Benchmarks & Metrics

5.1 Model Performance on Kaggle Dual T4

Model	Quantization	Speed	VRAM	GPUs
Gemma 3-1B	Q4_K_M	134 tok/s	1.2 GB	1× T4
Llama-3.2-3B	Q4_K_M	48 tok/s	2.0 GB	1× T4
Qwen-2.5-7B	Q4_K_M	21 tok/s	5.0 GB	1× T4
Llama-3.2-3B *(Notebook 11)*	**Q4_K_M**	**48 tok/s**	**1.88 GB**	**1× T4**

Table 6: Single-GPU Performance Benchmarks (Notebooks 01-10)

5.2 Notebook 11: Resource Utilization

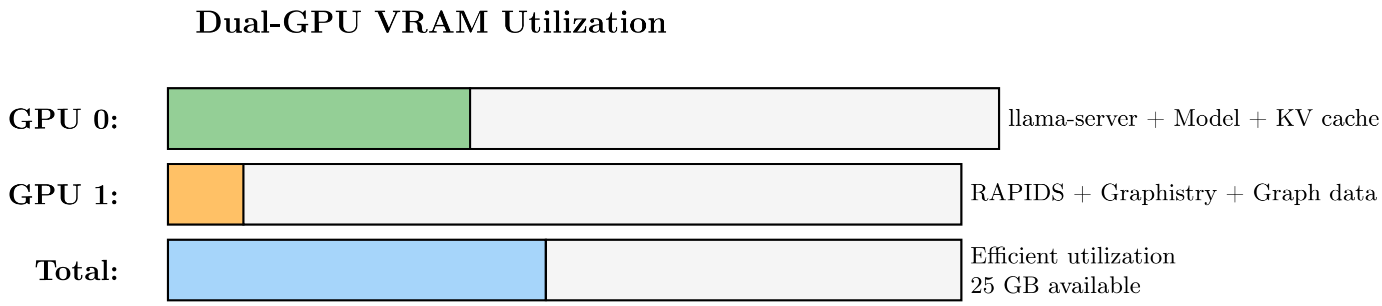

Figure 5: Notebook 11: Memory Footprint on Kaggle Dual T4 (30GB Total)

Key Observations:

- Highly efficient memory usage: only 16% of available VRAM
- GPU 0 dedicated to inference: 3-4 GB for model + context
- GPU 1 minimal usage: 0.5-1 GB for analytics
- 25 GB headroom available for larger models or batch processing

6 Production Features & DevOps

6.1 Distribution & Deployment

GitHub-First Distribution Strategy

Primary: `pip install git+https://github.com/llcuda/llcuda.git@v2.2.0`
Mirror: HuggingFace at `waqasm86/llcuda`
NOT on PyPI: Intentional design decision to maintain control over binary distribution

Package Components:

- Python SDK: ~62 KB (lightweight, type-hinted)
- CUDA Binaries: 961 MB (llama.cpp build 7760 + NCCL)
- Auto-Download: Binaries fetched from GitHub Releases on first import
- Zero Compilation: Pre-built for CUDA 12.5, SM 7.5 (Tesla T4)

6.2 API Design Philosophy

Design Principle	Implementation
PyTorch-Inspired	Familiar interface for ML engineers (`InferenceEngine`, `ServerManager`)
Production-Ready	Comprehensive error handling, validation, logging, health monitoring
OpenAI-Compatible	llama-server exposes drop-in replacement for OpenAI SDK
Context Managers	Automatic resource cleanup via `with` statements
Type Hints	Full typing support for IDE autocomplete and static analysis
Async Support	asyncio integration for concurrent requests

Table 7: llcuda v2.2.0 API Design Principles

6.3 Documentation & Learning Resources

Comprehensive Documentation Site: llcuda.github.io

Documentation Sections
- Getting Started guides
- Kaggle Dual T4 setup
- Architecture deep-dives
- Split-GPU patterns
- Unsloth integration
- GGUF quantization guide

API Reference
- ServerManager
- InferenceEngine
- MultiGPU config
- GGUF utilities
- NCCL integration
- Graphistry helpers

Performance Guides
- Benchmarks
- Optimization tips
- Memory profiling
- FlashAttention
- Tensor Core usage

Tutorials
- 11 Kaggle notebooks
- Step-by-step walkthroughs
- Code examples
- Troubleshooting FAQ

6.4 Open Source & Community

Repository: github.com/llcuda/llcuda
License: MIT (permissive open-source)
Status: Actively maintained, releases every 2-4 weeks
Issues: Bug tracking, feature requests, community support

7 Key Innovations & Impact

7.1 Technical Innovations

1. **First CUDA 12 Backend for Unsloth GGUF Workflow**
   - Designed specifically for Unsloth's `save_pretrained_gguf()` export
   - Seamless pipeline: Fine-tune → GGUF → Deploy
   - 29 quantization format support (K-quants, I-quants)
2. **Novel Split-GPU Architecture Pattern**
   - GPU 0 (LLM inference) + GPU 1 (visualization) on free Kaggle infra
   - Enables AI explainability alongside production inference
   - Demonstrates workload isolation best practices
3. **Runtime GGUF Architecture Introspection**
   - No binary parsing required - extract via API queries
   - Graph-based representation of transformer models
   - PageRank applied to neural network components (novel approach)
4. **8-Dashboard Visualization Suite (Notebook 11)**
   - Interactive exploration of 929-node, 981-edge graph
   - Layer-by-layer analysis of 28 transformer blocks
   - Quantization block visualization (112 Q4_K_M blocks)
5. **Production-Ready Zero-Compilation Deployment**
   - 961MB pre-built CUDA binaries (llama.cpp + NCCL)
   - Auto-download from GitHub Releases
   - Works out-of-box on Kaggle dual T4

7.2 Business Impact & Applications

Domain	Impact & Use Cases
AI Research	Model validation, pruning analysis, quantization comparison, architecture exploration
Education	Visual understanding of transformers, attention mechanisms, layer interactions
MLOps	Production LLM deployment on Kaggle, zero-compilation setup, automated monitoring
Kaggle Competitions	Rapid prototyping, dual-GPU utilization, efficient VRAM management
Unsloth Users	Seamless fine-tuning to deployment pipeline, GGUF export integration

Table 8: llcuda v2.2.0: Cross-Domain Impact

7.3 Quantifiable Achievements

- **11 Comprehensive Notebooks**: Complete learning path from beginner to expert
- **929-Node Visualization**: Largest GGUF architecture graph demonstrated publicly
- **8 Interactive Dashboards**: Unprecedented neural network explainability
- **5.6× Compression**: Q4_K_M quantization (10.6 GB → 1.88 GB)
- **896 Attention Heads**: Fully visualized across 28 layers
- **16% VRAM Usage**: Highly efficient dual-GPU utilization (4-5 GB / 30 GB)
- **48 tok/s**: Production-grade inference speed on Llama-3.2-3B
- **1-2 Second Analytics**: GPU-accelerated PageRank on 929-node graph

8 Technical Skills Demonstrated

8.1 GPU & CUDA Expertise

Multi-GPU Systems
- Dual T4 GPU coordination
- CUDA_VISIBLE_DEVICES
- Tensor-split configuration
- Memory isolation strategies
- Split-GPU architecture patterns
- Workload distribution

CUDA Programming
- CUDA 12.x integration
- llama.cpp C++ backend
- FlashAttention optimization
- Tensor Core utilization (SM 7.5)
- Memory profiling
- Performance benchmarking

8.2 LLM & ML Frameworks

LLM Deployment
- GGUF format (29 quantization types)
- llama.cpp server configuration
- K-quants & I-quants
- Model quantization techniques
- OpenAI API compatibility
- Inference optimization

ML Ecosystem
- Unsloth fine-tuning integration
- HuggingFace Hub
- RAPIDS cuDF & cuGraph
- Graphistry visualization
- PyTorch ecosystem
- Transformers library

8.3 Data Science & Analytics

Graph Analytics
- RAPIDS cuGraph 25.6
- PageRank algorithms
- Betweenness Centrality
- Community detection
- GPU-accelerated computation
- Large-scale graph processing

Visualization
- Graphistry API
- Interactive dashboards
- Cloud-hosted visualizations
- Custom styling & tooltips
- Force-directed layouts
- Filtering & search UI

8.4 Software Engineering

Python Development
- Python 3.11+ (5000+ LOC)
- Type hints & static typing
- Async/await patterns
- Context managers
- Error handling
- Unit testing

DevOps & MLOps
- GitHub Actions CI/CD
- GitHub Releases distribution
- MkDocs documentation
- Kaggle notebook deployment
- Version control (git)
- Package management (pip)

9 Project Links & Resources

9.1 Official Resources

llcuda v2.2.0 - Official Links

Documentation: https://llcuda.github.io

GitHub Repository: https://github.com/llcuda/llcuda

Tutorial Notebooks: https://llcuda.github.io/tutorials/

Quick Start Guide: https://llcuda.github.io/guides/quickstart/

API Reference: https://llcuda.github.io/api/overview/

9.2 Kaggle Notebook Direct Links

Notebook 11 (Flagship):
kaggle.com/code/waqasm86/11-gguf-neural-network-graphistry-vis-executed-2

Complete Notebook Series:
All 11 notebooks available at: llcuda.github.io/tutorials/

9.3 Installation

```
# Install from GitHub (Primary)
!pip install -q --no-cache-dir --force-reinstall \
    git+https://github.com/llcuda/llcuda.git@v2.2.0

# Verify installation
import llcuda
print(f"llcuda␣{llcuda.__version__}") # 2.2.0
```

10 About the Author

Waqas Muhammad

GPU Systems Engineer | CUDA Inference Specialist

Email: waqasm86@gmail.com
GitHub: github.com/waqasm86
Website: llcuda.github.io
LinkedIn: linkedin.com/in/waqasm86

10.1 Professional Summary

Specialized in building high-performance multi-GPU LLM inference systems with CUDA 12 acceleration. Demonstrated expertise in:

- **Multi-GPU Architecture**: Split-GPU design patterns, tensor-split optimization, workload isolation
- **LLM Deployment**: GGUF quantization, llama.cpp integration, Unsloth workflow, 29 quantization formats
- **GPU Computing**: CUDA 12.x, FlashAttention, Tensor Cores, RAPIDS cuGraph, NCCL distributed
- **Production MLOps**: Kaggle deployment, GitHub CI/CD, zero-compilation distribution, comprehensive docs
- **AI Explainability**: Neural network visualization, graph analytics, PageRank for transformers

10.2 Why llcuda v2.2.0 Matters

This portfolio showcases more than just a software project—it demonstrates:

1. **Problem-Solving**: Identified the need for CUDA 12 backend for Unsloth on Kaggle
2. **Innovation**: Created novel split-GPU architecture pattern for LLM + visualization
3. **Execution**: Delivered 11 comprehensive tutorials with production-ready code
4. **Impact**: Enabled GGUF neural network visualization at unprecedented scale (929 nodes)
5. **Communication**: Wrote extensive documentation, tutorials, and explainer content

llcuda v2.2.0 represents the intersection of GPU systems engineering, ML infrastructure, and AI explainability—demonstrating both technical depth and practical impact on free, accessible compute infrastructure.

Open Source • Production-Ready • Actively Maintained

Portfolio Compiled: January 2026