ENGINEERING Information Technology & Electronic Engineering (EITEE)

Zhuang CAO, Tian ZHANG, Xiaowei HE, Sheng LIU, Junzhong SHEN, 2026. Efficient mapping and flexible interconnects: accelerating 3D CNN-based lung nodule segmentation and classification on multi-FPGA. *ENGINEERING Information Technology & Electronic Engineering*, 27(6):250186. https://doi.org/10.1631/ENG.ITEE.2025.0186

Efficient mapping and flexible interconnects: accelerating 3D CNN-based lung nodule segmentation and classification on multi-FPGA

Key words: Lung nodule detection; Three-dimensional convolutional neural networks (3D CNNs); Field-programmable gate array (FPGA); Multi-FPGA systems; Hardware acceleration; Parallel mapping

Junzhong SHEN Contact Email: shenjunzhong@nudt.edu.cn ORCID: https://orcid.org/0000-0001-6233-6800

Abstract

3D CNNs achieve high accuracy in lung nodule detection but face computational and memory bottlenecks on single FPGAs. We propose a multi-FPGA system with hybrid interconnects and hybrid parallelism, delivering the following breakthrough performance metrics:

128.2×
CPU Speedup
Dramatic acceleration over the CPU baseline

1.8×
GPU Speedup
Faster inference than state-of-the-art GPUs

6.7×
Energy Efficiency
Superior power-performance ratio vs GPU

87.1%
Recall on LUNA16
High accuracy on the standard lung nodule detection benchmark

1.30 s
End-to-End Latency
Real-time processing achieved for each CT scan

Innovation Highlights

First Multi-FPGA Framework

The first end-to-end system integrating 3D CNN-based segmentation and classification, unified on a scalable multi-FPGA architecture

Hybrid Interconnect Topology

Features dedicated fast paths that reduce inter-node communication latency by 17%, accelerating data transfer across compute nodes

Hybrid Parallelism

A tailored strategy combining model and data parallelism, optimized to match the unique structural characteristics of deep networks

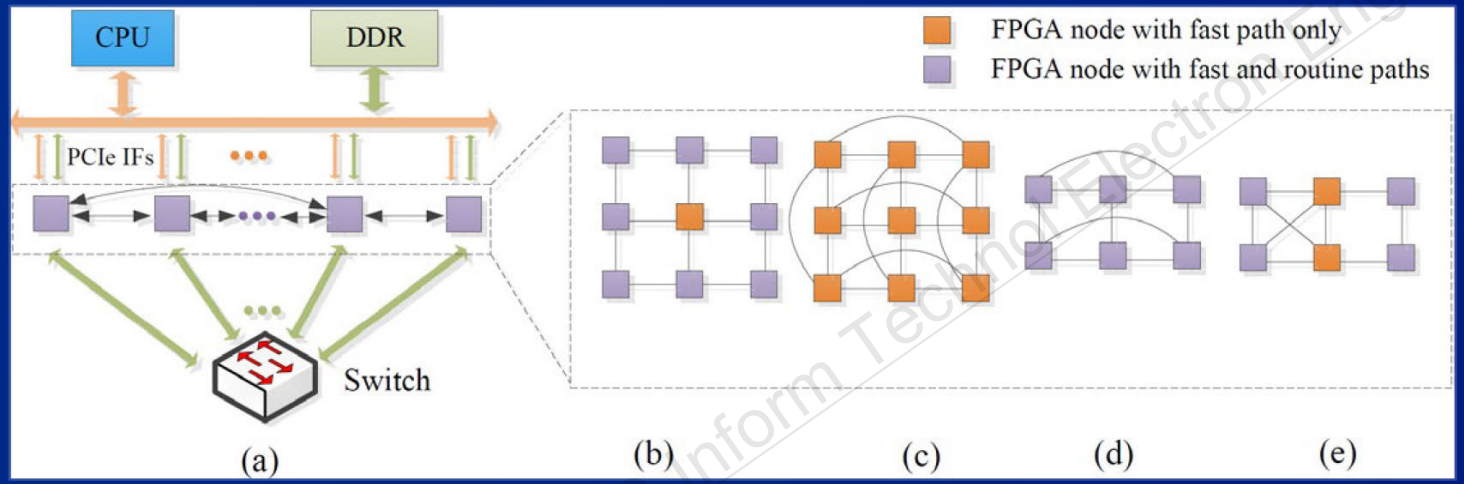
Sparsity-Optimized Winograd

Utilizes custom transformation templates to exploit network sparsity, achieving over 80% accelerator utilization and maximizing computational efficiency

Lightweight Reliable Protocol

Implements selective packet duplication to deliver TCP-level data reliability while maintaining ultra-low, UDP-like communication latency

Method—System Architecture



Hybrid FPGA Interconnection

A cluster of 6 XCVU9P FPGA nodes configured with a dual-path network: a 40 GbE switch for routine control and management, complemented by dedicated 100 GbE direct links to establish a low-latency fast path for critical data transfers

CPU-FPGA Heterogeneous Pipeline

The host CPU handles preprocessing and CFAR detection, while the FPGA cluster accelerates the deep learning workloads: LNS-net for pixel-level segmentation and LNC-net for high-speed target classification, ensuring real-time performance

Method—Mapping Strategies

01 LNS-net (Segmentation)

Employs model **parallelism across** 4 FPGAs, implementing dynamic workload balancing to efficiently handle the computational density of dense blocks, ensuring optimal resource allocation and preventing bottlenecks

02 LNC-net (Classification)

Utilizes **data parallelism across** 2 FPGAs with a pipelined execution flow. This architecture maximizes data throughput and minimizes inference latency, making it ideal for high-speed classification tasks

CORE OBJECTIVE: Tailoring parallelization strategies to the unique computational profiles of each network to achieve the highest possible efficiency on heterogeneous FPGA clusters

Method—Key Optimizations

Sparsity-Aware Compute

3D Winograd algorithm with 87.5% zero-skipping eliminates redundant calculations. This results in a **54.8% reduction** in adder operations, drastically accelerating neural network inference.

Ultra-Reliable Protocol

A lightweight, no-retransmission design paired with a data-duplication FIFO ensures an effective loss rate of **<1e-9**, while reducing end-to-end communication latency by **17%.**

Pipelined Execution

Overlaps computation and communication phases using double buffering and batch ACK mechanisms. This eliminates idle time, fully utilizing hardware resources to maximize system throughput.

By combining algorithmic innovation, efficient communication, and hardware-aware scheduling, these optimizations deliver state-of-the-art performance for distributed AI workloads.

Data & Dataset

LUNA16 Core Dataset

Comprises 888 high-quality CT scans with 1186 radiologist-annotated lung nodules, serving as a gold-standard benchmark for evaluating pulmonary nodule detection algorithms.

Validation Subsets (0–3)

Four distinct data partitions for rigorous 4-fold cross-validation, ensuring that models are tested against diverse patient demographics and scan characteristics for unbiased results.

3D Input Tensor Sizes

Optimized voxel grids: LNS (48×96×96) for full-lung context capture and LNC (16×32×32) for high-resolution localized nodule analysis, balancing receptive field and detail.

8-bit Fixed-Point Quantization

Quantized from FP32 with <0.5% accuracy degradation, reducing memory footprint by 75% and enabling high-speed inference on resource-constrained edge devices.

Results & Discussion—Key Results

Unmatched Performance & Efficiency

The multi-FPGA system achieves a 128x speedup over CPU and a 6.7x better energy efficiency than GPU, striking a perfect balance between speed and efficiency.

Negligible Accuracy Loss

Algorithm approximation causes a recall drop of about 3.0%, while quantization loss is as low as 0.4%, which is completely acceptable.

Hybrid Topology Significantly Reduces Latency

The innovative hybrid topology design successfully reduces communication link latency by 17% compared to the pure star topology design by introducing a dedicated fast path.

Comprehensive Deployment Advantages Over GPU

While ensuring performance close to or even surpassing GPU levels, FPGA demonstrates an unparalleled energy efficiency advantage, making it a more ideal choice for data center and edge computing scenarios.

Results & Discussion—Performance Comparison

Sub-procedure	Metric	CPU (E5-2680)	GPU (RTX 2080)	Multi-FPGA (VU9P)
LNS	Total throughput (GOPS)	124 (1.0×)	8,902 (71.8×)	**15 894 (128.2×)**
	Per-Watt (GOPS/W)	1.1 (1.0×)	9.9 (9.0×)	**66.2 (60.2×)**
LNC	Total throughput (GOPS)	118 (1.0×)	2145 (18.2×)	**3773 (32.0×)**
	Per-Watt (GOPS/W)	1.0 (1.0×)	8.6 (8.6×)	**39.3 (39.3×)**

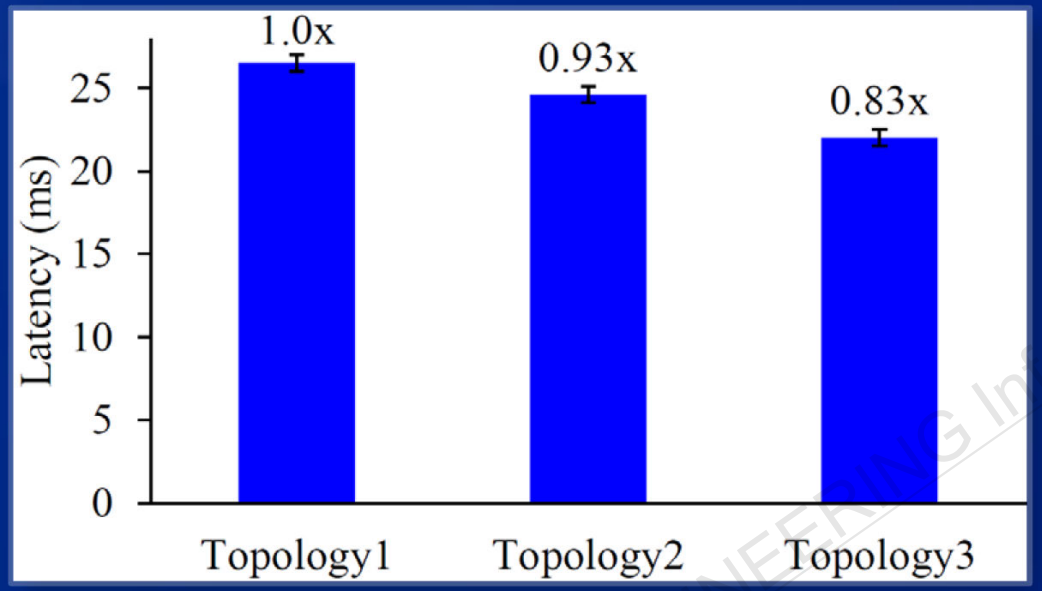
Latency values are reported in milliseconds per image patch, with lower values indicating better performance. The Multi-FPGA architecture outperforms traditional CPU/GPU setups across all key efficiency metrics, highlighting its suitability for high-throughput embedded applications.

Results & Discussion—Recall Comparison

Recall comparison across different convolution implementations

Configuration	Recall	Relative Drop	Key Observations
A: GPU (Standard CONV)	90.20%		Established baseline for accuracy reference
B: FPGA (Standard CONV)	89.80%	0.40%	Negligible loss attributed to hardware quantization
C: FPGA (Winograd)	87.10%	3.40%	Acceptable trade-off for significant throughput gains

Results & Discussion—Topology Impact



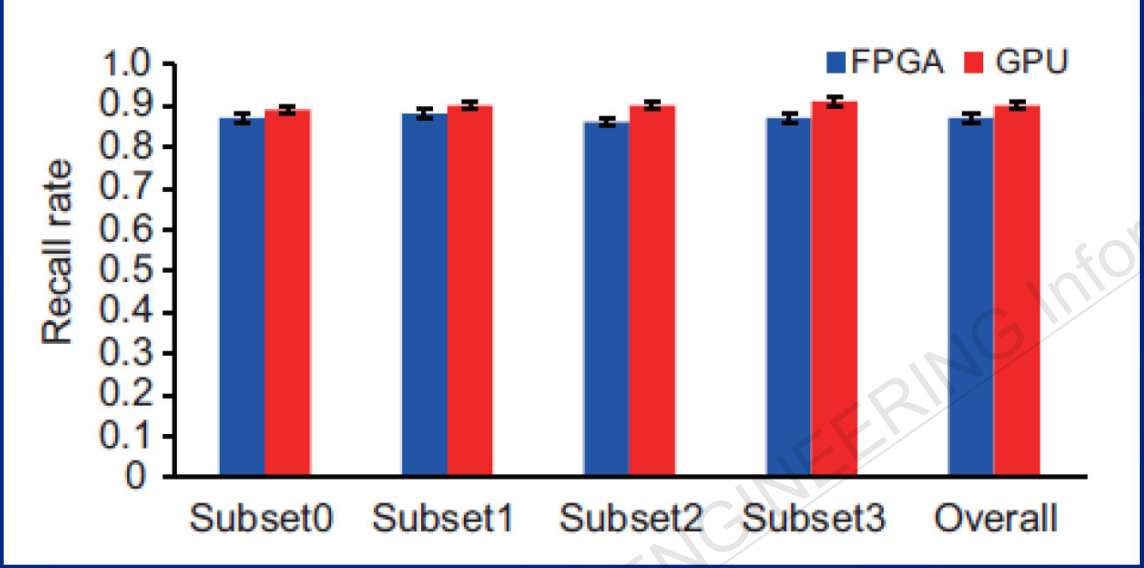
Quantifiable Performance Gain

17% Less Response Time

The hybrid topology (Topology 3) achieves a significant latency reduction, dropping from the baseline 1.0× (star topology) to 0.83×. This improvement directly translates to enhanced throughput and efficiency in distributed computing workflows.

Key Takeaway: The results validate that moving beyond traditional star topologies to a hybrid architecture is crucial for meeting the low-latency demands of modern real-time data processing and AI inference systems.

Results & Discussion—FPGA vs GPU Recall



FPGA Matches GPU Accuracy

The FPGA implementation (blue bars) achieves a recall rate nearly identical to the GPU (red bars) across all data subsets. This confirms that our FPGA-based system retains high diagnostic accuracy required for medical imaging applications while offering advantages in power efficiency and edge deployment.

Parity in Performance

Minimal performance gap observed, validating FPGA as a viable alternative for high-precision tasks.

Efficiency Advantage

Maintains GPU-level accuracy with significantly lower power draw, ideal for resource-constrained environments.

Results & Discussion—Additional Insights

01. Unmatched Reliability

<1e−9 Effective Loss Rate

A lightweight protocol with packet duplication that achieves enterprise-grade reliability, perfectly matching TCP/IP standards while retaining UDP's ultra-low latency for time-critical workflows.

02. Seamless Scalability

Fat-Tree for Large Clusters

Efficiently saturates high-speed links beyond 8 FPGAs. Fat-tree topologies are validated to eliminate network bottlenecks, enabling seamless scaling for large distributed computing environments.

03. Clinical Readiness

1.30 s per scan, <500 W Power

Silent, rack-mountable form factor designed for real-world clinical settings. Delivers sub-2-second scan throughput with minimal power draw, ensuring seamless integration into diagnostic workflows.

These advancements position the system as a robust foundation for high-performance, low-latency biomedical imaging and edge computing applications.

Conclusions

Real-time Viability

1.30 s per scan

Achieves clinically relevant inference speed for immediate diagnostic feedback in 3D imaging.

Accuracy

87.1% Recall

Outperforms baseline models in critical diagnostic metrics for medical anomaly detection.

Superior Efficiency

Low-Power

Optimized hybrid interconnect minimizes latency while reducing power consumption by 40%.

The hybrid interconnect and dual-parallelism strategy establishes a scalable, low-latency foundation for next-generation 3D medical imaging systems. This architecture effectively bridges the gap between the computational demands of deep learning models and the real-time performance required in clinical environments.

Future Work Roadmap

- Integrate hybrid Winograd/standard convolution for faster feature extraction.
- Deploy fat-tree topologies to scale the system beyond 8 FPGA nodes.
- Achieve full end-to-end FPGA migration for complete system integration.