Deepak Inugala

Senior HPC Engineer | GPU Cluster Architecture | AI / LLM Infrastructure | High-Speed Networking | International Delivery

Email: deepak.1990@hotmail.com
Phone: +971 50 494 5921
Location: Abu Dhabi, UAE
LinkedIn: linkedin.com/in/deepak-inugala

NVIDIA H100/H200	MIG & NCCL	InfiniBand	Air-Gapped	6 Countries	99.99%
GPU Cluster Expert	HPC Optimization	High-Speed Fabric	Sovereign HPC	International Delivery	Cluster Uptime

PROFESSIONAL SUMMARY

Senior HPC Engineer with 10+ years of infrastructure engineering experience, specializing in designing, provisioning, and operating enterprise GPU clusters and high-performance computing environments for AI, LLM inference, and distributed training workloads. Deep hands-on expertise with NVIDIA H100 and H200 GPU servers, Multi-Instance GPU (MIG) partitioning, InfiniBand / NVLink high-speed fabric, NCCL collective communication tuning, and Kubernetes-native GPU orchestration (GPU Operator, DCGM Exporter, KubeRay). Proven track record at Group 42 (G42) building production HPC infrastructure for sovereign AI platforms across GOVINT, OSINT, and Smart Nation verticals — including fully air-gapped deployments. International HPC delivery across 6 countries (on-site in Kazakhstan, Angola, Bahrain, UAE; remote for Maldives and Ethiopia).

HPC & GPU INFRASTRUCTURE SKILLS

GPU Hardware	NVIDIA H100 (SXM5/PCIe), H200, A100, A30; bare-metal provisioning — BIOS/UEFI, RAID, iDRAC/iLO, GPU health validation; Dell PowerEdge, HP ProLiant, Huawei server platforms
MIG & GPU Slicing	Multi-Instance GPU (MIG) configuration and profile management (1g.10gb – 7g.80gb); GPU Operator MIG Manager for K8s-native automated slice allocation; time-slicing for dev/test
High-Speed Fabric	InfiniBand (HDR 200Gb/s, NDR 400Gb/s), NVLink 4.0 (900GB/s), RoCE v2; Mellanox/NVIDIA ConnectX NIC management; ibstat, ib_write_bw benchmarking; OpenSM subnet manager
NCCL & Distributed	NCCL tuning (GDR, QPS, buffer sizes, ring/tree algorithms); GPU Direct RDMA (NCCL_NET_GDR_LEVEL=5); AllReduce benchmarking via nccl-tests; multi-node distributed training optimization
CUDA & Driver Stack	NVIDIA driver lifecycle (offline install, version pinning); CUDA, cuDNN, NCCL library compatibility; GPU firmware updates via offline bundles; DCGM diagnostic suites
K8s GPU Orchestration	NVIDIA GPU Operator (driver daemonset, container toolkit, device plugin, DCGM Exporter, MIG Manager); KubeRay (RayCluster CRD); nvidia.com/gpu and nvidia.com/mig-* resource limits; GPU node taints/tolerations
LLM Serving	vLLM (PagedAttention, tensor parallelism, continuous batching, AWQ/GPTQ quantization); NVIDIA Triton Inference Server (TensorRT, ONNX, dynamic batching); LiteLLM gateway; ONNX Runtime CUDAExecutionProvider
GPU Observability	DCGM Exporter (GPU utilization, memory, temperature, NVLink bandwidth, ECC errors); Prometheus + Grafana GPU dashboards; Nsight Systems (nsys) profiling; nvidia-smi, nvtop; MTTD < 3 min via DCGM alerts
Storage for HPC	Ceph (Rook-Ceph operator), NFS high-throughput mounts, Longhorn, MinIO / S3-compatible storage for model artifacts and checkpoints; PV/PVC design for large model weight volumes
Air-Gapped HPC	Full offline GPU cluster bootstrap: NVIDIA driver offline packages, MLNX_OFED tarball, container image tarballs, Helm chart bundles; Harbor with nvcr.io/docker.io mirrors; HF_HUB_OFFLINE=1; model weight pipeline
Cloud & Kubernetes	AKS (GPU node pools), EKS, RKE2, Rancher; Terraform (GPU infra provisioning), Ansible (OS hardening, NVIDIA driver install, containerd); Azure (Key Vault, Defender), G42 Cloud
Linux & Networking	Ubuntu 22.04/24.04, RHEL 8/9; kernel tuning for GPU (hugepages, NUMA, vm.swappiness=0, inotify); CIS hardening via Ansible; Palo Alto Firewalls, VLAN, NIC bonding (LACP), S2S/P2S VPN

WORK EXPERIENCE

Senior HPC Engineer | GPU Infrastructure | SRE | Group 42 (G42)

07/2019 – Present | Abu Dhabi, UAE | On-Site: Kazakhstan, Angola, Bahrain | Remote: Maldives, Ethiopia

International HPC Delivery — On-Site & Remote (6 Countries)

- Deployed and commissioned G42 AI and HPC platforms at client sites across Kazakhstan, Angola, and Bahrain — sole technical authority for GPU infrastructure, networking, storage, and security; additionally supported Maldives and Ethiopia remotely via secure VPN. Managed all pre-deployment readiness: data center vendor coordination, GPU server rack-and-stack, high-speed networking commissioning, storage validation, and client UAT — travelling approximately 40% of each month; led knowledge-transfer workshops and formally handed over fully operationalized systems with runbooks and SLA documentation.
- Delivered HPC and AI infrastructure concurrently across GOVINT (classified air-gapped GPU clusters for government intelligence), OSINT (GPU-accelerated NLP inference and large-scale data processing), and Smart Nation projects (real-time AI inference and city-scale analytics pipelines).

GPU Cluster Design, Provisioning & Operations

- Designed and provisioned enterprise GPU clusters from bare metal — racked NVIDIA H100 (SXM5) and H200 servers, validated dual-PDU power and cooling capacity, configured BIOS/UEFI for GPU workload optimization (NUMA, hugepages), and performed per-GPU health validation using nvidia-smi and DCGM diagnostic suites before cluster integration; managed complete NVIDIA driver and firmware lifecycle including offline update bundles via iDRAC across production GPU fleets.
- Configured Multi-Instance GPU (MIG) profiles on H100 nodes using GPU Operator MIG Manager — defined mixed MIG strategies (e.g. 2x 3g.40gb + 1x 1g.10gb per GPU) to serve concurrent LLM inference workloads of different model sizes; managed profile updates via Kubernetes ConfigMaps with zero-disruption rolling restarts; deployed GPU Operator on both AKS and bare-metal RKE2 managing driver daemonset, container toolkit, device plugin, DCGM Exporter, and MIG Manager as a unified lifecycle.

InfiniBand / NVLink Fabric & NCCL Tuning

- Administered HDR InfiniBand (200Gb/s) fabric across multi-node GPU clusters — configured Mellanox ConnectX-6 NIC drivers (MLNX_OFED offline install), managed OpenSM subnet manager for IB fabric topology, validated inter-node bandwidth with ib_write_bw achieving >190Gb/s sustained throughput, and diagnosed fabric issues using ibstat and infiniband-diags; monitored NVLink 4.0 topology within 8-GPU H100 nodes via DCGM metric DCGM_FI_DEV_NVLINK_BANDWIDTH_TOTAL.
- Tuned NCCL collective communications for distributed LLM training and inference — enabled GPU Direct RDMA (NCCL_NET_GDR_LEVEL=5) to bypass CPU in data path, set NCCL_IB_QPS_PER_CONNECTION=4 for better IB bandwidth utilization, configured NCCL_BUFFSIZE=8388608 for large AllReduce operations; validated near-linear scaling across 16-GPU multi-node setups using nccl-tests benchmarks.

LLM Inference Infrastructure

- Built production vLLM serving infrastructure on GPU clusters — deployed large-scale LLMs (7B-72B parameters) with tensor parallelism, AWQ 4-bit quantization (reducing 72B model from 144GB to 36GB), continuous batching, and prefix caching; achieved P99 TTFT under 200ms at 1,000+ concurrent requests. Deployed NVIDIA Triton Inference Server for embedding and encoder models with TensorRT optimization (2-3x speedup) and dynamic batching via gRPC endpoints for low-latency RAG pipelines.
- Configured AWQ and GPTQ quantization pipelines for VRAM optimization; validated quantized model quality against MMLU and HumanEval benchmarks before production promotion; managed HuggingFace model downloads, offline packaging, and containerization for air-gapped deployments.

GPU Observability, Performance & Air-Gapped HPC

- Integrated DCGM Exporter with Prometheus and Grafana — built real-time GPU health dashboards covering utilization, memory, temperature, NVLink bandwidth, and ECC errors; configured Alertmanager rules for GPU hang detection, thermal throttling, and ECC double-bit errors with immediate pod eviction; implemented Kubernetes HPA on DCGM metrics for vLLM auto-scaling; reduced GPU incident MTTR by 40%. Conducted GPU profiling with NVIDIA Nsight Systems to identify CUDA kernel bottlenecks, recovering 30% throughput regression in production inference code.
- Designed and executed fully air-gapped GPU cluster deployments for classified GOVINT and sovereign client environments — built complete offline artifact pipelines (NVIDIA driver .run packages, MLNX_OFED tarball, GPU Operator Helm bundle, container image tarballs, HuggingFace model packages with pre-applied quantization); configured Harbor with nvcr.io, docker.io mirrors; set HF_HUB_OFFLINE=1 and TRANSFORMERS_OFFLINE=1 on all inference pods; maintained model artifact versioning in internal MinIO.

MLOps Compute Infrastructure

- Built and operated end-to-end MLOps compute infrastructure on GPU clusters — provisioned Kubeflow Pipelines with MinIO and MySQL backend; deployed KubeRay operator for distributed fine-tuning jobs with auto-scaling GPU

worker groups; integrated MLflow experiment tracking and model registry with automated promotion gates feeding ArgoCD-triggered vLLM deployment updates — reducing model release cycles from weeks to hours.

Cloud Engineer | First Abu Dhabi Bank (FAB)

02/2018 – 07/2019 | Abu Dhabi, UAE

- Cloud Infrastructure & Big Data Administration: completed FGB-NBAD bank merger IT integration with zero major incidents; designed and managed AWS architectures (EC2, EKS, RDS, S3, Lambda); installed and administered Cloudera CDH clusters (HDFS NameNode HA, YARN, Hive, HBase, Spark) with Kerberos authentication and Ranger authorization; deployed multi-node Elasticsearch clusters with ILM policies and Kibana dashboards.
- Vulnerability Management & Monitoring: owned full patching lifecycle — CVE triage, Ansible-based OS patching, post-patch scan validation; implemented centralized monitoring via CloudWatch, Prometheus, and Grafana; led incident response reducing repeat incidents by 35%.

Linux Engineer | HCL InfoSystems Limited

04/2015 – 01/2018 | Abu Dhabi, UAE

- Large-Scale Linux & Data Center Operations: managed 3,000+ physical Linux servers (RHEL, CentOS) — OS provisioning, Ansible configuration management, hardware fault diagnosis (iDRAC/iLO), and firmware lifecycle management across Dell, HP, and IBM server platforms in enterprise data centers.
- Network, Firewall & Migration: configured Cisco/IBM switches (VLAN management, STP/RSTP, LACP); administered firewall rule sets, DMZ segmentation, and NAT rules; led on-premises-to-cloud migration of Linux infrastructure and Cloudera Hadoop clusters with under 2 hours downtime and 0% data loss.

CERTIFICATIONS

Microsoft Certified: Azure Administrator Associate (AZ-104)
Valid: 02/2026 - 07/2027

Microsoft Certified: Azure Solutions Architect Expert (AZ-305)
Valid: 07/2025 - 07/2027

AWS Certified Solutions Architect - Associate
Valid: 07/2022 - 07/2025 | ID: PFR7HTQ25EFQQS9T

Certified Kubernetes Administrator (CKA)
Valid: 09/2021 - 09/2023 | Renewal in Progress

G42 Cloud Certified Engineer
Valid: 10/2021 - 12/2024 | ID: G42C/SVD/CRT/0475

Red Hat Certified Engineer (RHCE)
Valid: 02/2015 - 02/2018 | ID: 150-012-904

EDUCATION

Master of Business Administration (MBA)
Jawaharlal Nehru Technology University | 2012 - 2015

Bachelor of Technology (B.Tech)
Jawaharlal Nehru Technology University | 2008 - 2012

LANGUAGES & SOFT SKILLS

Languages

English & Hindi (Professional) | Telugu (Native) | German (Elementary)

Soft Skills

Cross-functional Collaboration | Stakeholder Communication | Problem Solving | Rapid Learning | Adaptability