



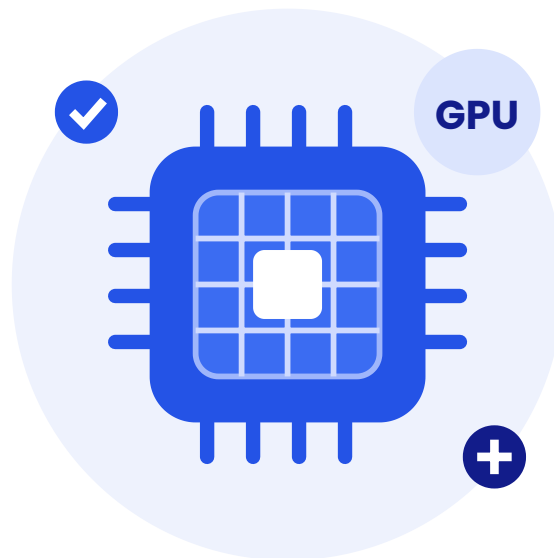
CareerByteCode
Learning Made simple

CareerByteCode Admissions

+91 63817 54092

support@careerbytecode.in

www.careerbytecode.in



CAREERBYTECODE PROGRAM

Enterprise AI Factory: Architecture, Engineering & Operations

An 8-weekend live cohort on designing, building and operating enterprise GPU clusters: GPU and rack-scale systems, InfiniBand and RoCE fabrics, lossless networking, power and liquid cooling, storage, BCM provisioning, Kubernetes and Run:ai orchestration, and production troubleshooting, with labs on a live HPC cluster.

**Build the AI Factory.
Run it in production.**

Instructors: Aravind Kumar TS & Madhumathy U

Why CareerByteCode?

LEARNING MADE SIMPLE

CareerByteCode gives you many platforms to work with, with real-time hands-on practice across many use cases. In this program you do not just study AI infrastructure, you operate it: on real GPUs, real fabrics and real schedulers, alongside a global community built to help you grow and get hired.



A Live HPC Cluster, Not Slides

Every learner gets personal SSH access to a multi-GPU HPC cluster with NVIDIA accelerators, ConnectX SuperNICs and high-speed fabrics, available 24/7 for practice.



Real-Time, Hands-On

Around half of the program is keyboard-level CLI work: perfest RDMA benchmarks, device plugins, Slurm scripts and Kubernetes manifests.



Real Enterprise Use Cases

Reference designs, exploded BOMs, client acceptance protocols and site surveys taken from real AI Factory deployments, so your skills map to real projects.



Mentorship & Growth

Direct mentorship from practising AI consultants, resume reviews for AI infrastructure roles, and an active community for ongoing networking.

CareerByteCode
Learning Made simple

What You Get

FROM CAREERBYTECODE

This is an execution-first program. You leave with real cluster experience, a design portfolio and a community, not just a certificate.

✓ 8 Weekends, 16 Live Sessions

Mentor-led sessions every Saturday and Sunday, 9:00 PM to 10:30 PM IST.

✓ 23 In-Depth Modules

From GPU silicon and NVL72 racks to fabrics, facilities, orchestration and runbooks.

✓ 10 Hands-On HPC Cluster Labs

Real hardware labs with step-by-step terminal runbooks and pre-tested templates.

✓ Dedicated Cluster Access

Personal SSH credentials with CUDA, Slurm, Docker, Kubernetes and DCGM pre-installed.

✓ Lifetime Session Recordings

HD recordings, slide decks, network diagrams, cheat sheets and a Git repo of every runbook.

✓ NVIDIA Certification Support

Practice question bank and strategy sessions for NCA-AIIO, NCP-AIN and NCP-AII.

✓ Career & Interview Prep

Resume reviews and architecture interview rounds for AI infrastructure, HPC and cloud roles.

✓ Verifiable Certificate

A CareerByteCode certificate signed by Sangeetha, Director, verifiable on careerbytecode.in.

CareerByteCode
Learning Made simple

The CareerByteCode Ecosystem

SEVEN PILLARS OF GROWTH

Every learner plugs into one connected ecosystem built for tech professionals: to learn, build, get hired and earn.

- 1 Learning Platform**
100-day challenge tracks, domain challenges and career roadmaps, free and production-grade.
- 2 Tech Consulting**
Corporate training, upskilling, POCs and architecture guidance for enterprises.
- 3 Community & Events**
Hackathons, webinars, workshops and LinkedIn Lives reaching 242K+ learners.
- 4 Freelancing & Consulting**
Client projects, paid mentorship and 300+ earning opportunities.
- 5 SaaS Partnerships**
Product demos, beta testing and developer workshops with leading SaaS companies.
- 6 Hiring & Recruitment**
A talent network connecting skilled professionals with companies worldwide.
- 7 Revenue Sharing**
Industry-leading 55% to 80% revenue share for trainers and service providers.

Earn while you grow: 55% to 80% revenue share

Industry-leading revenue share for trainers, service providers and vendors, with 300+ earning opportunities across training delivery, consulting, course sales, reselling, speaking and mentorship.

CareerByteCode
Learning Made simple

Program Format & Schedule

LIVE WEEKEND COHORT

DURATION

8 Weekends

FORMAT

Live Cohort (mentor-led)

SESSIONS

16 live, 1.5 hrs each

MODULES

23 across 4 Parts

STACK

NVIDIA GPUs, InfiniBand, RoCE, Kubernetes

LEVEL

Intermediate (Linux and networking basics)

Your Weekend Rhythm

- 2 live mentor-led sessions every weekend (Saturday and Sunday)
- Guided lab plus a main lab or design exercise every weekend
- Weekday self-paced lab hours on the HPC cluster, with mentor office hours
- Live Q&A, troubleshooting drills and doubt-clearing

Upcoming Batch Schedule

Start Date	Batch Type	Timing
Rolling batches	Weekend Batch (Sat and Sun)	9:00 PM to 10:30 PM IST, 1.5 hr per session

New batches start regularly due to high demand. Message us on WhatsApp for the next start date.

How You Learn



CareerByteCode
Learning Made simple

Who Should Join

BUILT FOR WORKING ENGINEERS

For network, data centre, cloud, storage, DevOps and Linux systems engineers moving into AI infrastructure roles, and anyone preparing for NVIDIA AI infrastructure and networking certifications.



Cloud & DevOps Engineers

Moving into high-performance AI infrastructure, GPU containerization and Kubernetes GPU scheduling.



Data Centre, Systems & Network Admins

Managing servers, switching and storage, and ready to master liquid-cooled racks, InfiniBand and automated bring-up.



Platform & AI Infrastructure Architects

Technical leads who design, size (BOM) and procure enterprise AI clusters and DGX SuperPODs.



AI/ML Engineers & MLOps Professionals

Practitioners who want to understand HBM limits, communication patterns and how to maximize MFU.

Prerequisites

LINUX

Comfortable with the CLI, file systems and bash scripting

NETWORKING

IP addressing, subnets, VLANs, routing and the OSI model

CONTAINERS (HELPFUL)

Awareness of Docker and Kubernetes is a bonus

No prior InfiniBand, AI, deep learning or Python experience is required. InfiniBand is taught from the ground up and all distributed benchmark scripts are pre-configured.

CareerByteCode
Learning Made simple

Table of Contents

ENTERPRISE AI FACTORY PROGRAM

WEEKEND 01 • AI WORKLOADS, PARALLELISM & COLLECTIVES

01	AI Fundamentals for Infrastructure Engineers	08
02	AI Workloads and Network Traffic	08
03	AI Parallelism Strategies	08
04	Collective Communications & NCCL	08

WEEKEND 03 • SCALE-OUT AI FABRICS & TOPOLOGIES

07	Scale-Out AI Networks	10
08	AI Network Topologies	10
13	Load Balancing & Dynamic Routing	10

WEEKEND 05 • AI CLUSTER & DATA CENTRE DESIGN

14	Training vs Inference Networks	12
15	AI Data Centre & GPU Cluster Design	12
16	Ultra Ethernet	12

WEEKEND 07 • PROVISIONING, SOFTWARE STACK & ORCHESTRATION

20	NVIDIA Software Stack & AI Ecosystem	14
21	Cluster Management & Workload Scheduling	14

PROGRAM RESOURCES

•	Hands-on HPC Cluster Labs	16
•	NVIDIA Certification Prep	17
•	Your Mentors	18
•	Final Outcomes	19
•	Career Path & How to Join	20

WEEKEND 02 • GPU SYSTEMS & RACK-SCALE ARCHITECTURE

05	GPU and Compute System Architecture	09
06	Scale-Up Networking	09
+	Rack Facilities & Bill of Materials	09

WEEKEND 04 • RDMA & LOSSLESS NETWORKING

09	RDMA Fundamentals & Zero-Copy Architecture	11
10	RDMA Transport Technologies	11
11	Lossless Ethernet for AI	11
12	Congestion Control Mechanisms	11

WEEKEND 06 • POWER, COOLING & AI STORAGE

17	Power Infrastructure for AI Data Centres	13
18	Advanced Thermal Management & Cooling	13
19	Storage & Checkpointing for AI	13

WEEKEND 08 • OBSERVABILITY, RESILIENCY & ACCEPTANCE

22	GPU Monitoring and Observability	15
23	Troubleshooting, Validation & Acceptance	15

01 Module 1: AI Fundamentals for Infrastructure Engineers

Why AI workloads run on GPUs, and why the network becomes the cluster.

Topics covered in the **AI Fundamentals for Infrastructure Engineers** module include:

- AI vs ML vs Deep Learning for Systems Engineers
- Why GPUs and Accelerators Beat CPUs for AI
- Model Parameter Scaling
- Compute Bounds vs Memory Bandwidth Limits
- Why the Fabric Becomes Mission-Critical at Scale
- MFU: Measuring Real Compute Efficiency

02 Module 2: AI Workloads and Network Traffic

The traffic profile an AI fabric must carry.

Topics covered in the **AI Workloads and Network Traffic** module include:

- North-South vs East-West Traffic
- Elephant Flows and Synchronized Phases
- Microburst Dynamics
- Incast: Many-to-One and Buffer Exhaustion
- Forward Pass, Backward Pass and All-Reduce
- Why AI Overturns Enterprise Network Assumptions

03 Module 3: AI Parallelism Strategies

How model parallelism shapes cluster traffic.

Topics covered in the **AI Parallelism Strategies** module include:

- Data Parallelism (DP) and FSDP
- Tensor Parallelism (TP) Inside the Node
- Pipeline Parallelism (PP) and Pipeline Bubbles
- 3D and Hybrid Parallelism
- Aligning Models with Network Topology
- Profiling PyTorch and TensorFlow Jobs

04 Module 4: Collective Communications & NCCL

The collectives every distributed training job depends on.

Topics covered in the **Collective Communications & NCCL** module include:

- Broadcast, Gather, Scatter, Reduce
- All-Reduce, All-Gather, Reduce-Scatter
- Ring vs Tree vs Double-Binary Tree
- NCCL Environment Tuning
- Synchronization Barriers and Tail Latency
- How Collectives Drive Fabric Design

GUIDED LAB

Profile a distributed PyTorch job

MAIN LAB

Lab 2: Collectives in Action (NCCL Ring vs Tree)

05 Module 5: GPU and Compute System Architecture

From silicon to rack: Hopper, Blackwell and the GB300 NVL72.

Topics covered in the **GPU and Compute System Architecture** module include:

- CPU, GPU, TPU, DPU and xPU Roles
- Hopper (H100/H200) vs Blackwell (B200/GB200/GB300)
- Tensor Cores, Transformer Engine, FP4/FP8
- HBM3e vs GDDR vs SRAM and NUMA Locality
- PCIe Gen5/Gen6, Root Complexes and Switches
- DGX and HGX Baseboards and SXM Modules
- ConnectX SuperNICs and BlueField DPUs
- Scale-Up vs Scale-Out Systems

06 Module 6: Scale-Up Networking

Turning up to 72 GPUs into one memory-coherent domain.

Topics covered in the **Scale-Up Networking** module include:

- NVLink 5 Generations and Signaling
- NVSwitch 5 Fabric ASICs
- Connecting 8, 16 and 72 GPUs
- Blind-Mate Copper Backplane vs Optics
- UALink and SUE-T Open Standards
- Scale-Up Design Principles

AI FACTORY Rack Facilities & Bill of Materials

AI Factory deep dive: what a 132 kW rack is made of.

Topics covered in the **Rack Facilities & Bill of Materials** module include:

- 120 to 132 kW Power Shelves
- 415V/480V 3-Phase AC to 48V DC Busbars
- Direct-to-Chip Liquid Cooling Loops
- CDUs and Leak Detection
- Exploded BOM: Compute and Switch Trays
- OOB Switches and Cabling for Proposals

GUIDED LAB

Lab 1: Meet the GPU Node (nvidia-smi, lspci, topology)

MAIN LAB

Size an exploded BOM for a GB300 NVL72 rack

07 Module 7: Scale-Out AI Networks

InfiniBand and Spectrum-X Ethernet fabrics for GPU clusters.

Topics covered in the **Scale-Out AI Networks** module include:

- Quantum-2 and Quantum-X800 InfiniBand
- Subnet Manager (OpenSM)
- Spectrum-X Ethernet and SuperNICs
- Credit-Based Flow Control vs PFC/ECN
- Fat-Tree and Non-Blocking Clos
- Oversubscription and Bisection Bandwidth
- In-Network Computing with NVIDIA SHARP
- UFM Telemetry, BER Audits, Cable Health

08 Module 8: AI Network Topologies

Choosing the right topology for the cluster you are building.

Topics covered in the **AI Network Topologies** module include:

- Multi-Tier Clos and High-Radix Switching
- Rail-Optimized 2-Tier vs 3-Tier Fat-Tree
- Managing the 3-Hop to 5-Hop Transition
- Dragonfly and Dragonfly+
- Slimfly and Diameter Reduction
- 1D, 2D and 3D Torus Interconnects
- Cost, Cabling and Resilience Trade-Offs
- OOB Management and North-South Ingress

13 Module 13: Load Balancing & Dynamic Routing

Keeping elephant flows from colliding.

Topics covered in the **Load Balancing & Dynamic Routing** module include:

- Why ECMP Fails with Elephant Flows
- Flowlet Switching
- Packet Spraying
- Adaptive Routing and Queue Inspection
- RDMA Packet Reordering and Reassembly
- Maximizing GPU Utilization

GUIDED LAB

Compare topologies on cost, hops and cabling

MAIN LAB

Read a UFM fabric telemetry and BER report

Learning Made simple

09 Module 9: RDMA Fundamentals & Zero-Copy Architecture

How GPUs move data without touching the CPU.

Topics covered in the **RDMA Fundamentals & Zero-Copy Architecture** module include:

- Kernel Bypass and Zero-Copy Transfers
- Queue Pairs, Work and Completion Queues
- Memory Registration and Protection Domains
- RDMA Verbs: Read, Write, Atomic
- GPUDirect RDMA Across Nodes

10 Module 10: RDMA Transport Technologies

Picking the right transport for production.

Topics covered in the **RDMA Transport Technologies** module include:

- Native InfiniBand vs RoCEv1 vs RoCEv2
- iWARP over TCP
- Packet Formats and Encapsulation
- Latency and Throughput Benchmarks
- Choosing a Production Transport

11 Module 11: Lossless Ethernet for AI

Why one dropped packet stalls a training step.

Topics covered in the **Lossless Ethernet for AI** module include:

- The Cost of Drops in RDMA
- PFC (802.1Qbb) and Headroom Buffers
- Head-of-Line Blocking and PFC Deadlocks
- DCB and DCBX
- ETS (802.1Qaz) Bandwidth Allocation

12 Module 12: Congestion Control Mechanisms

Tuning the fabric before congestion spreads.

Topics covered in the **Congestion Control Mechanisms** module include:

- ECN Marking and RED Thresholds
- DCQCN: CNP and Rate Control
- TIMELY, HPCC and Swift
- In-band Network Telemetry (INT)
- Production Tuning Recipes

GUIDED LAB

Lab 3: RDMA Basics & Validation (ib_write_bw)

MAIN LAB

Lab 4: Lossless Ethernet Tuning (PFC, ECN, DCQCN)

14 Module 14: Training vs Inference Networks

Designing separate networks for separate jobs.

Topics covered in the **Training vs Inference Networks** module include:

- Training Traffic: Bursty All-Reduce
- Inference Traffic: Low-Latency Streaming
- Frontend vs Backend vs OOB Networks
- Dedicated Low-Jitter Inference Design

15 Module 15: AI Data Centre & GPU Cluster Design

From one pod to a SuperPOD.

Topics covered in the **AI Data Centre & GPU Cluster Design** module include:

- Rail-Optimized vs Rail-Unified Designs
- Single-Rail vs Multi-Rail NIC Mapping
- GPU Pods and Scalable Units (SUs)
- NVIDIA SuperPOD Reference Architectures
- Spine-Leaf Cabling and Optics
- Failure Domains and Blast Radius
- Digital Twin Design with NVIDIA DSX Air
- Maintenance and CapEx Trade-Offs

16 Module 16: Ultra Ethernet

Where Ethernet for AI is heading.

Topics covered in the **Ultra Ethernet** module include:

- Ultra Ethernet Consortium (UEC) Vision
- Limits of Classic 802.3 Ethernet
- Multi-Pathing and Packet Spraying
- Out-of-Order Delivery, Selective Retransmit
- UEC vs RoCEv2 vs InfiniBand
- Ethernet Roadmap in HPC

GUIDED LAB

Map NICs to rails for an 8-GPU node

MAIN LAB

Lab 5: Fabric Design Studio (32-node rail-optimized pod)

CareerByteCode
Learning Made simple

17 Module 17: Power Infrastructure for AI Data Centres

Feeding 40 to 132 kW racks reliably.

Topics covered in the **Power Infrastructure for AI Data Centres** module include:

- 8 to 15 kW Racks vs 40 to 132 kW Racks
- 3-Phase 415V/480V and Overhead Busways
- Smart PDUs and Power Capping
- N+1, 2N and UPS Battery Sizing
- PUE and Carbon Efficiency

18 Module 18: Advanced Thermal Management & Cooling

Why air runs out above 40 kW per rack.

Topics covered in the **Advanced Thermal Management & Cooling** module include:

- Air Cooling Limits
- Rear-Door Heat Exchangers (RDHx)
- Direct-to-Chip (DTC) Cold Plates
- Single-Phase vs Two-Phase Immersion
- CDUs, Manifolds and Facility Loops
- Leak Detection and Fluid Chemistry

19 Module 19: Storage & Checkpointing for AI

Keeping checkpoints off the training fabric.

Topics covered in the **Storage & Checkpointing for AI** module include:

- Storage Across the AI Lifecycle
- HBM vs Local NVMe vs Shared Storage
- Lustre, GPFS (Storage Scale), WEKA, VAST
- Checkpoint Bursts and Recovery Time
- GPUDirect Storage (GDS) and NVMe-oF
- Dedicated Storage Rails vs HoL Blocking

GUIDED LAB

Pre-deployment site survey: power and CDU audit

MAIN LAB

Lab 6: Storage Benchmarking with FIO and gdscheck

CareerByteCode
Learning Made simple

20 Module 20: NVIDIA Software Stack & AI Ecosystem

Everything that runs between the GPU and the model.

Topics covered in the **NVIDIA Software Stack & AI Ecosystem** module include:

- GPU Drivers, Kernel Modules, CUDA
- Container Toolkit and NGC Containers
- MIG Partitioning vs vGPU and vWS
- NVIDIA AI Enterprise: TensorRT, RAPIDS, NeMo
- Serving with Triton and NVIDIA NIM

21 Module 21: Cluster Management & Workload Scheduling

Bring-up with BCM, then schedule with Slurm, Kubernetes and Run:ai.

Topics covered in the **Cluster Management & Workload Scheduling** module include:

- Base Command Manager: Head Nodes, PXE Boot
- Golden Images, OFED, CUDA, Zero Drift
- Slurm Partitions, QoS and GPU GRES
- GPU Operator and Network Operator
- Multus CNI for RDMA in Pods
- Run:ai Fractional GPUs and KAI Scheduler
- Fair-Share Queues and Preemption
- Slurm vs Kubernetes and Gang Scheduling

GUIDED LAB

Lab 7: Build the Software Stack and slice a GPU with MIG

MAIN LAB

Lab 8: Schedule GPU Jobs on Slurm and Kubernetes

CareerByteCode
Learning Made simple

22 Module 22: GPU Monitoring and Observability

Seeing problems before a multi-day job stalls.

Topics covered in the **GPU Monitoring and Observability** module include:

- nvidia-smi Query Loops and Telemetry
- DCGM Health Checks and Diagnostic Levels
- DCGM Exporter, Prometheus and Grafana
- InfiniBand Symbol Errors (OpenSM/UFM)
- Ethernet Pause Frame Counters
- BMC and Redfish Telemetry

23 Module 23: Troubleshooting, Validation & Acceptance

The runbooks that keep production clusters healthy.

Topics covered in the **Troubleshooting, Validation & Acceptance** module include:

- NCCL Tests and the perfctest Suite
- XID Errors (31, 45, 62, 79) and ECC Faults
- Thermal Throttling and PCIe Dropouts
- MTBF and Stragglers at 1,000+ GPUs
- AHR Node Cordoning, AJR Job Recovery
- 3-Phase Client Acceptance Protocol
- 24-Hour HPL Burn-In and DCGM Level 3
- Runbooks for Cordon, Drain and RMA

GUIDED LAB**Lab 9: Monitor the Cluster with DCGM and Prometheus****MAIN LAB****Lab 10: Break-Fix Scenarios on injected faults**

CareerByteCode
Learning Made simple

Hands-on HPC Cluster Labs

10 REAL-HARDWARE LABS

Every part ends with practical labs on our dedicated HPC cluster. You practise on real enterprise GPU hardware, high-speed fabrics and orchestration tools.

Lab	Title	Hands-on Practical Exercises	Modules
Lab 1	Meet the GPU Node	Inspect GPUs, memory hierarchy, PCIe root complexes and NVLink topology using <code>nvidia-smi</code> , <code>lspci -tv</code> and topology matrix commands.	1, 5, 6
Lab 2	Collectives in Action	Run NCCL benchmarks (<code>all_reduce_perf</code>), compare Ring vs Tree efficiency, and launch a distributed PyTorch benchmark across GPUs.	3, 4
Lab 3	RDMA Basics & Validation	Query RDMA devices with <code>ibv_devinfo</code> , configure Queue Pairs, and run <code>ib_write_bw</code> and <code>ib_read_lat</code> between nodes.	9, 10
Lab 4	Lossless Ethernet Tuning	Configure PFC, ECN and DCQCN on switches and SuperNICs, and observe flow control under simulated congestion.	11, 12
Lab 5	Fabric Design Studio	Design a rail-optimized fabric for a 32-node GPU pod: switch ports, spine-leaf oversubscription and cable tables.	7, 8, 13, 15
Lab 6	Storage Benchmarking	Benchmark parallel file system throughput with FIO, time checkpoint reads and writes, and configure GPUDirect Storage (<code>gdscheck</code>).	19
Lab 7	Build the Software Stack	Install drivers, CUDA and the NVIDIA Container Toolkit, partition a GPU into MIG slices, and run jobs in NGC containers.	20
Lab 8	Schedule GPU Jobs	Write sbatch scripts requesting multi-GPU resources, deploy the GPU Operator on Kubernetes, and schedule inference pods.	21
Lab 9	Monitor the Cluster	Deploy DCGM Exporter, configure Prometheus collection, and build Grafana views for GPU utilization, power and temperature.	22
Lab 10	Break-Fix Scenarios	Diagnose live injected faults: failed links, thermal and power throttling, driver mismatches and real XID errors.	23



Dedicated Cluster Environment

Individual SSH credentials to a multi-GPU cluster with NVIDIA accelerators, ConnectX SuperNICs and high-speed fabrics.



Step-by-Step Lab Runbooks

Complete terminal guides, pre-tested configuration templates and benchmark datasets for every lab.

NVIDIA Certification Prep

INDUSTRY CREDENTIALS

The program prepares you for three official NVIDIA certifications, from associate to professional level.

Certification Track	Level	Exam Format	Mapped Parts
AI Infrastructure and Operations NCA-AIIO	Associate	50 multiple choice questions, 60 minutes	Parts A, C, D
AI Networking NCP-AIN	Professional	70 to 75 questions, 2 hours, optional lab exercises	Parts B, C
AI Infrastructure NCP-AII	Professional	Advanced architecture and operational assessment	Parts B, C, D

Certification Support Bank

- Curated practice questions matching NVIDIA exam objectives
- Deep-dive question walkthroughs and exam strategy sessions
- Mock exams and revision in Weekend 08
- Guidance on test booking and syllabus updates

Exam formats are referenced from official NVIDIA certification pages (checked October 2026); always confirm the latest format on NVIDIA's certification site. This is an independent training program by CareerByteCode and is not affiliated with or endorsed by NVIDIA Corporation.

CareerByteCode
Learning Made simple

Your Mentors

LEARN FROM PRACTISING AI CONSULTANTS

You learn directly from consultants who build and operate AI infrastructure, with 1-on-1 mentorship throughout the cohort.

AK

Aravind Kumar TS

AI CONSULTANT

15 years of experience in IT infrastructure, cloud computing and AI infrastructure, with expertise in NVIDIA GPU platforms, Kubernetes, MLOps and scalable AI environments.

linkedin.com/in/professionalsiaas

MU

Madhumathy U

AI CONSULTANT

Application Support Executive at Elektroxen (R&D), supporting NXP microcontroller boards and building user interfaces in C for embedded systems. Works with 8051, Arduino UNO, LPC804 and LPC845 ARM controllers and protocols such as UART, SPI, I2C and CAN, and has contributed to numerous IoT projects that integrate hardware and software.

linkedin.com/in/madhuinsights

CareerByteCode
Learning Made simple

Final Outcomes

WHAT YOU WALK AWAY WITH

By the end of the 8 weekends, you will be able to:

- ✓ Architect high-density GPU racks: thermal (CDU liquid cooling) and electrical (132 kW) anatomy of GB300 NVL72 and DGX systems
- ✓ Design rail-optimized 2-tier and 3-tier InfiniBand and RoCE fat-tree fabrics, with UFM telemetry and adaptive routing
- ✓ Engineer lossless networking: RDMA, RoCEv2, PFC, ECN and DCQCN, configured and troubleshot
- ✓ Automate cluster provisioning with Base Command Manager: golden images, OFED, CUDA and DCGM diagnostics
- ✓ Orchestrate GPUs with Slurm, the GPU Operator, Multus CNI and Run:ai fractional GPUs and fair-share queues
- ✓ Optimize storage with GPUDirect Storage and parallel file systems, isolated from collective traffic
- ✓ Run client acceptance and QoS testing: site surveys, XID triage and multi-node NCCL sweeps with zero stragglers

Tools & Technologies Covered



CareerByteCode
Learning Made simple

Your Career Path

FROM LEARNING TO EARNING

This program is designed to take you from your first module to your first AI infrastructure role.

1 Master the 23 Modules

Go from GPU silicon and rack facilities through fabrics, storage, orchestration and production runbooks.

2 Practise on a Real HPC Cluster

Complete 10 hands-on labs and break-fix drills on real GPUs, SuperNICs and high-speed fabrics.

3 Get Certification Ready

Prepare for NVIDIA NCA-AIIO, NCP-AIN and NCP-AII with mock exams and strategy sessions.

4 Step Into AI Infrastructure Roles

Resume reviews and design interview prep for AI infrastructure, HPC and cloud roles.



Join the Upcoming Batch

Batch timings, registration details, syllabus updates and announcements are shared in our WhatsApp community. Scan the code to join and reserve your seat for the next cohort.

chat.whatsapp.com/E8U8kuSOHJX323D1DjMkvX

Or message us on WhatsApp at +91 63817 54092.

CareerByteCode
Learning Made simple



CareerByteCode
Learning Made simple

THANK YOU

CareerByteCode Admissions • +91 63817 54092
support@careerbytecode.in • www.careerbytecode.in

CareerByteCode is a brand of Cloudncloud Technologies Pvt Ltd, Chennai