MIDDLEWARE 2024

WoSC 10

10th International Workshop on Serverless Computing

Advancing Serverless Computing for Scalable AI Model Inference: Challenges and Opportunities

Li Wang

Yankai Jiang

Ningfang Mi

Northeastern University

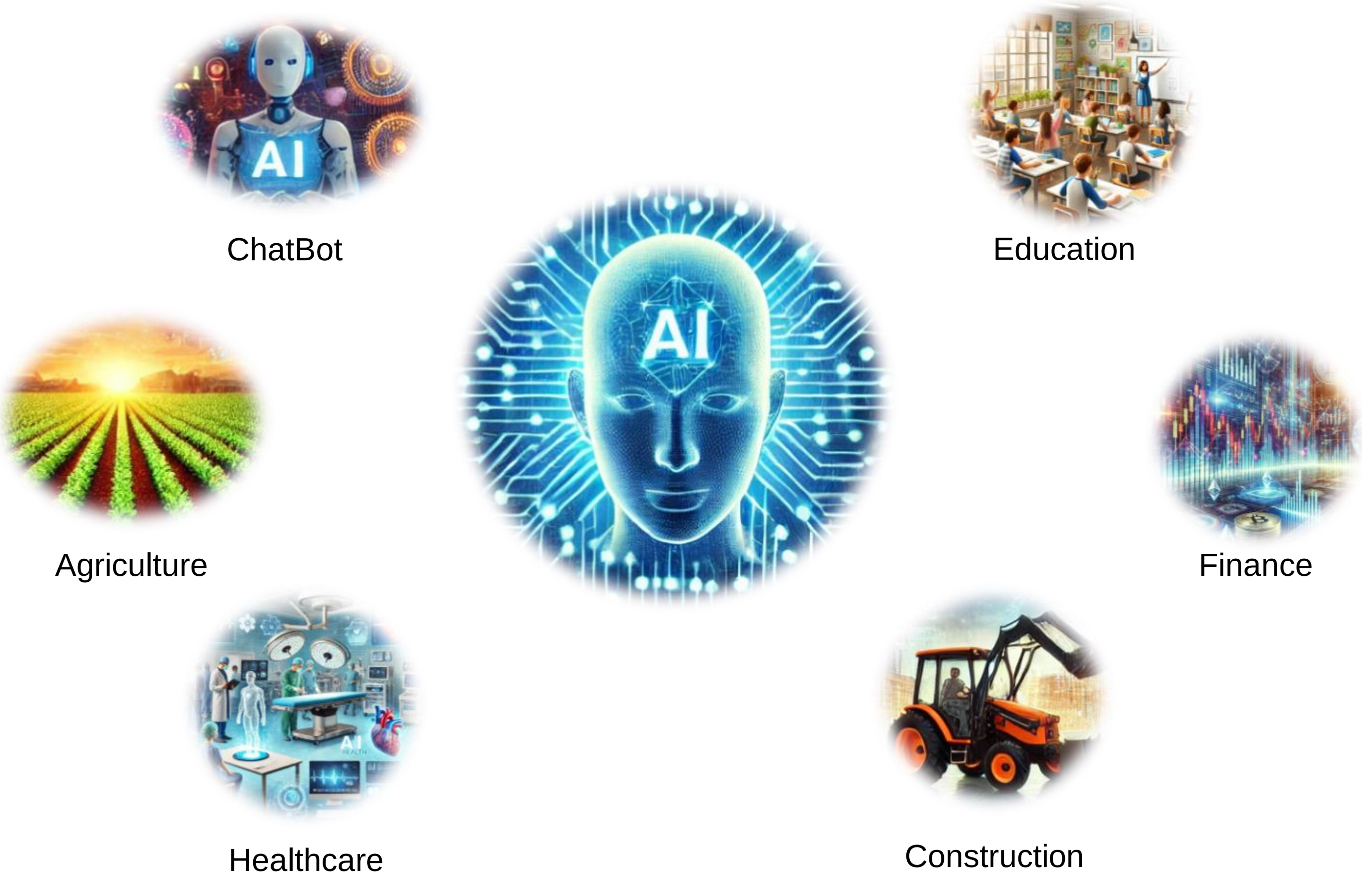
2024.12.02

Outline

- Motivations
- Contributions
- Challenges
- Optimal Strategies
- Emerging Research Fields
- Insights
- Conclusions

Motivations

AI is everywhere ……

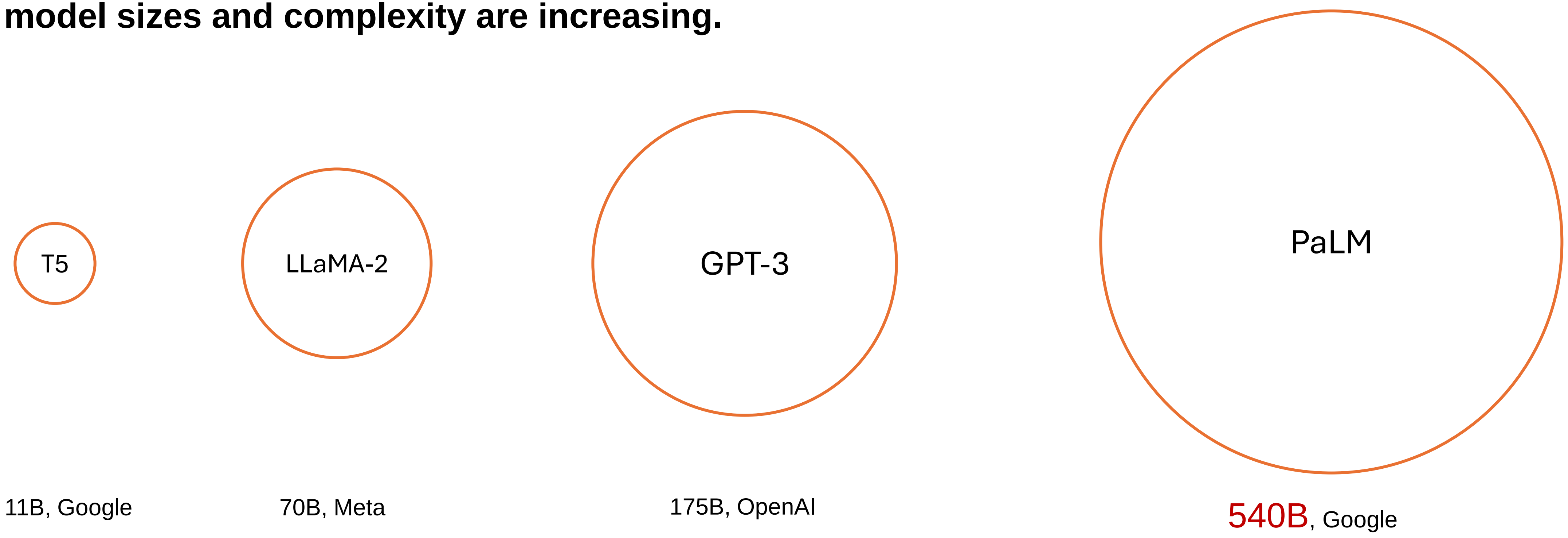


Images sources: Internet

Northeastern University

Motivations

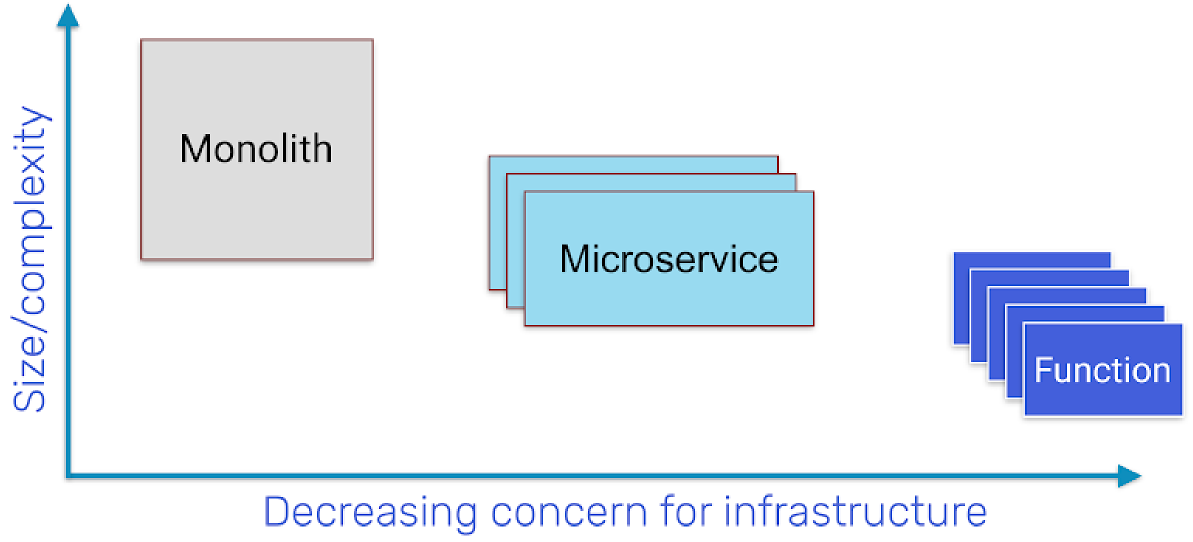
AI model sizes and complexity are increasing.



Numbers of parameters in pre-trained LLMs (having a size larger than 10B) in recent years.

Motivations

Cloud computing paradigm is changing from IaaS to FaaS.

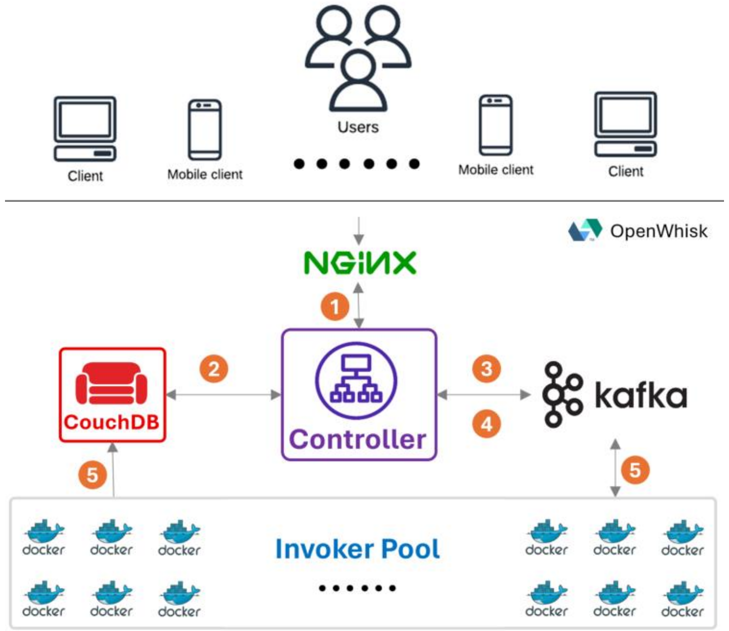


Cloud computing service models.
Source: Peter Mell, Timothy Grance, The NIST Definition of Cloud Computing, National Institute of Standards and Technology (NIST)

Northeastern University

Motivations

Emerging serverless paradigm with elastic resource scaling and pay-per-use billing model.



A workflow of function invocation with serverless paradigm.
Source: Wang et al., Uncovering The Impact of Bursty Workloads on System Performance in Serverless Computing, ISNCC, 2024

Northeastern University

Contributions

Both academic and industry are making efforts to understand and optimize the deployment of AI model inference systems with serverless paradigm.

83 Existing Works

- Conferences
- Journals
- arXiv
- Theses
- Innovations in Industry

Northeastern University

Contributions

Both academic and industry are making efforts to understand and optimize the deployment of AI model inference systems with serverless paradigm.

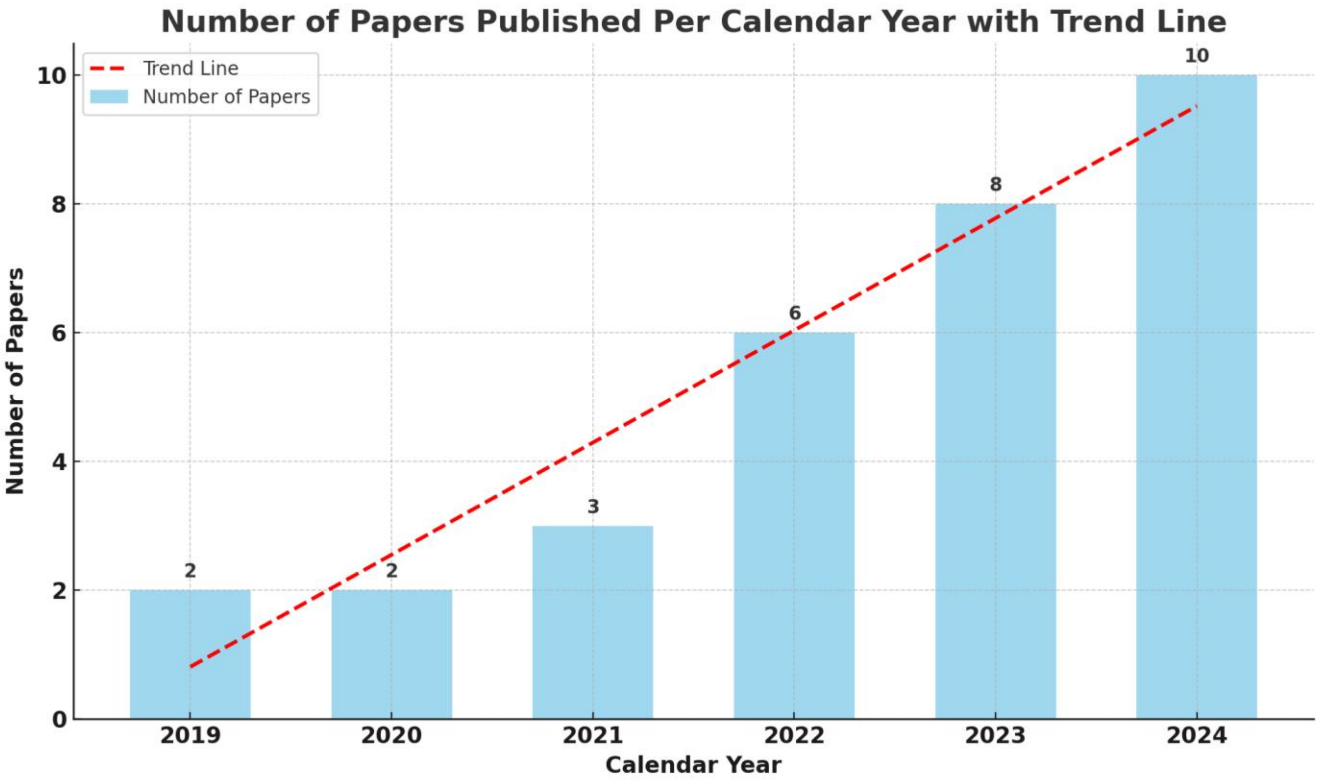
83 Existing Works →

- Publication Date
- Conference Venue
- Paper Category
- Keywords
- ……
- ……
- ……

→ **31 Selected Works**

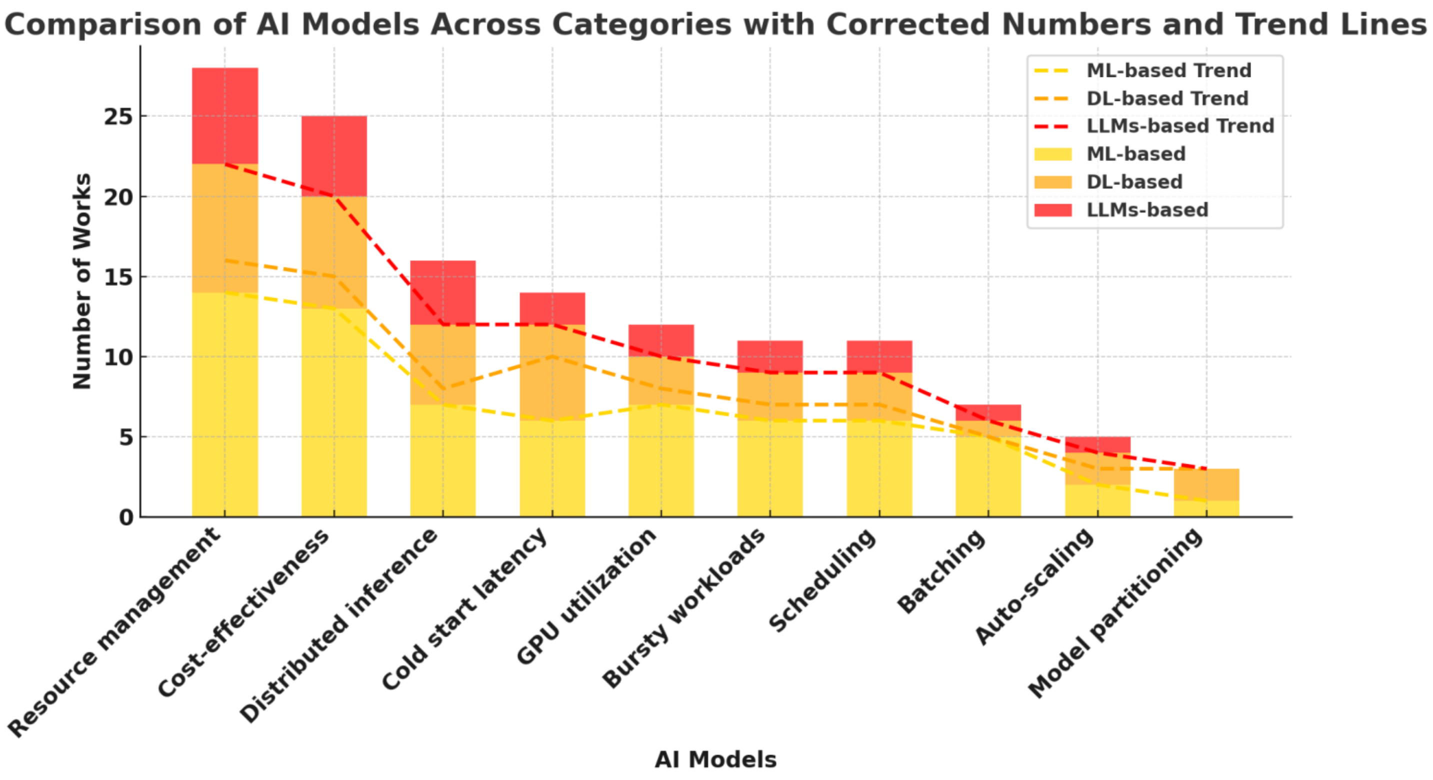
Contributions

The selected 31 works show an increasing trend across year (2019.01 ~ 2024.09).



Contributions

From the 31 selected works, we classify them into ML-, DL-, LLMs-based inference. Subsequently, we further divide these works into 10 subcategories for detailed analysis.



Contributions

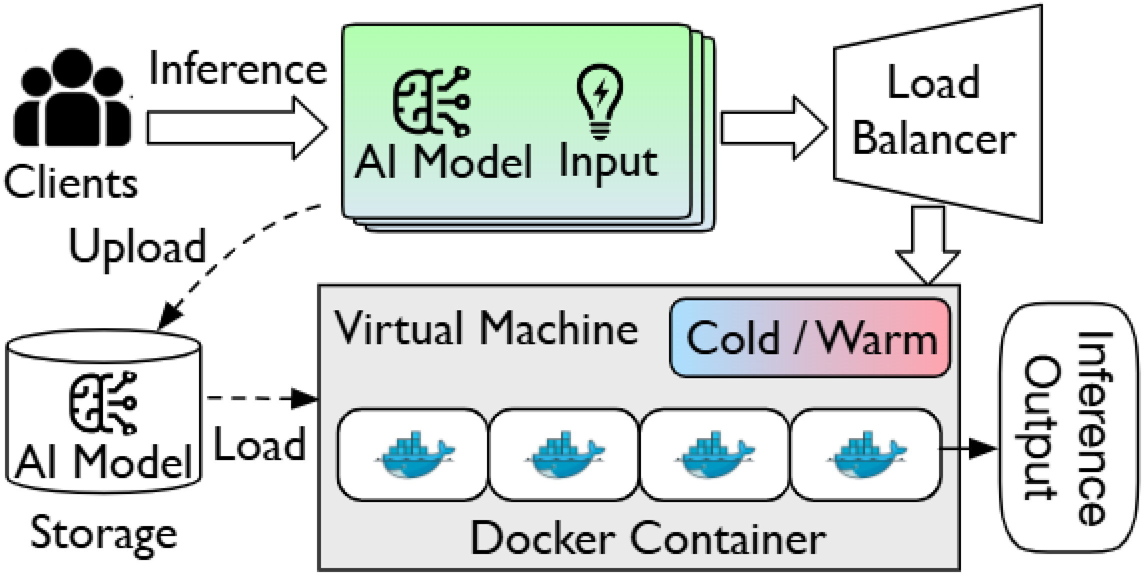
From the 31 selected works, we classify them into ML-, DL-, LLMs-based inference. Subsequently, we further divide these works into 10 subcategories for detailed analysis.

AI Models	ML-based	DL-based	LLMs-based
Resource management	[1, 3, 6, 7, 10, 12, 13, 15, 19, 22, 23, 25, 28, 31]	[5, 8, 11, 16, 20, 24, 27, 30]	[2, 9, 14, 17, 21, 26]
Cost-effectiveness	[1, 6, 7, 10, 12, 13, 15, 19, 22, 23, 25, 28, 31]	[5, 8, 11, 20, 24, 27, 30]	[2, 9, 14, 21, 26]
Distributed inference	[15, 19, 22, 25, 27, 28, 31]	[5, 10, 11, 16, 27]	[9, 14, 17, 26]
Cold start latency	[1, 7, 12, 13, 25, 28]	[7, 8, 16, 20, 23, 30]	[9, 26]
GPU utilization	[1, 6, 7, 12, 13, 19, 31]	[11, 16, 20]	[9, 14]
Bursty workloads	[1, 6, 12, 13, 25, 31]	[5, 16, 24]	[9, 26]
Scheduling	[1, 6, 12, 13, 28, 31]	[5, 20, 24]	[9, 26]
Batching	[1, 6, 12, 28, 31]	[5]	[26]
Auto-scaling	[22, 31]	[5, 20]	[14]
Model partitioning	[10]	[8, 30]	N/A

Statistics of 10 trending topics in ML-, DL-, LLMs-based inference.

Background

Deploy AI model inference systems with serverless paradigm on the cloud.



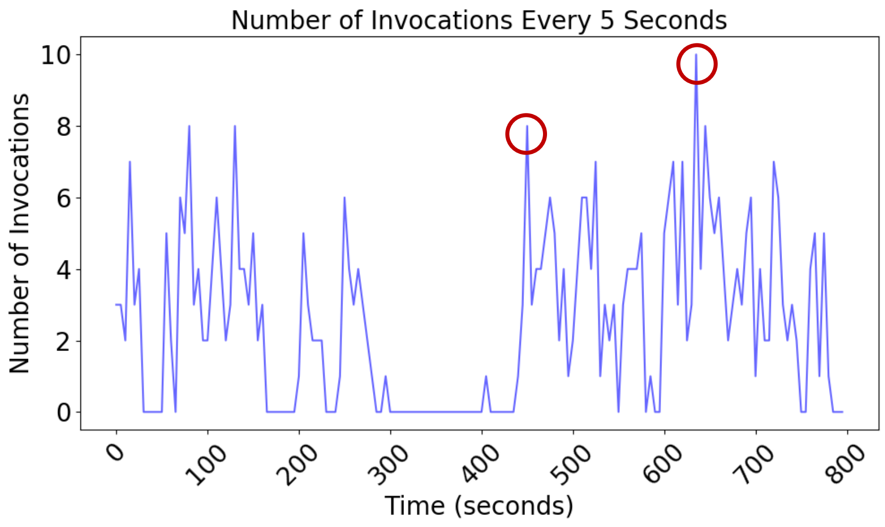
A workflow of serverless inference process.

Challenges

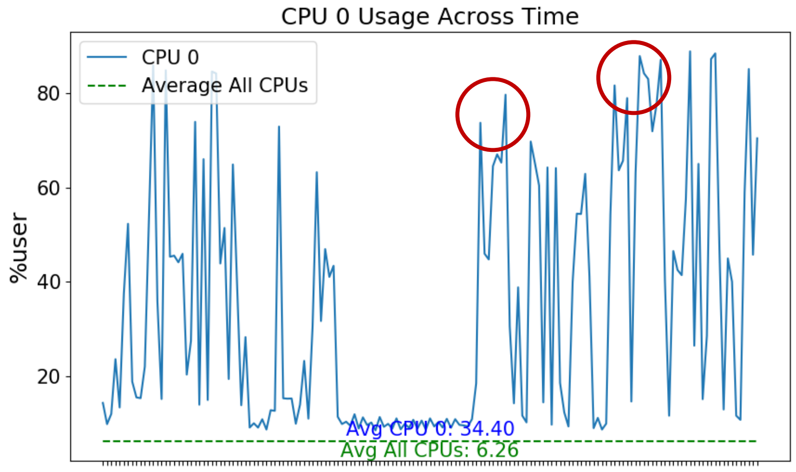
Challenges in scalable AI model inference deployment with serverless paradigm.

Challenges

Dynamic workloads with spikes (bursty workloads).



An example dynamic workload.



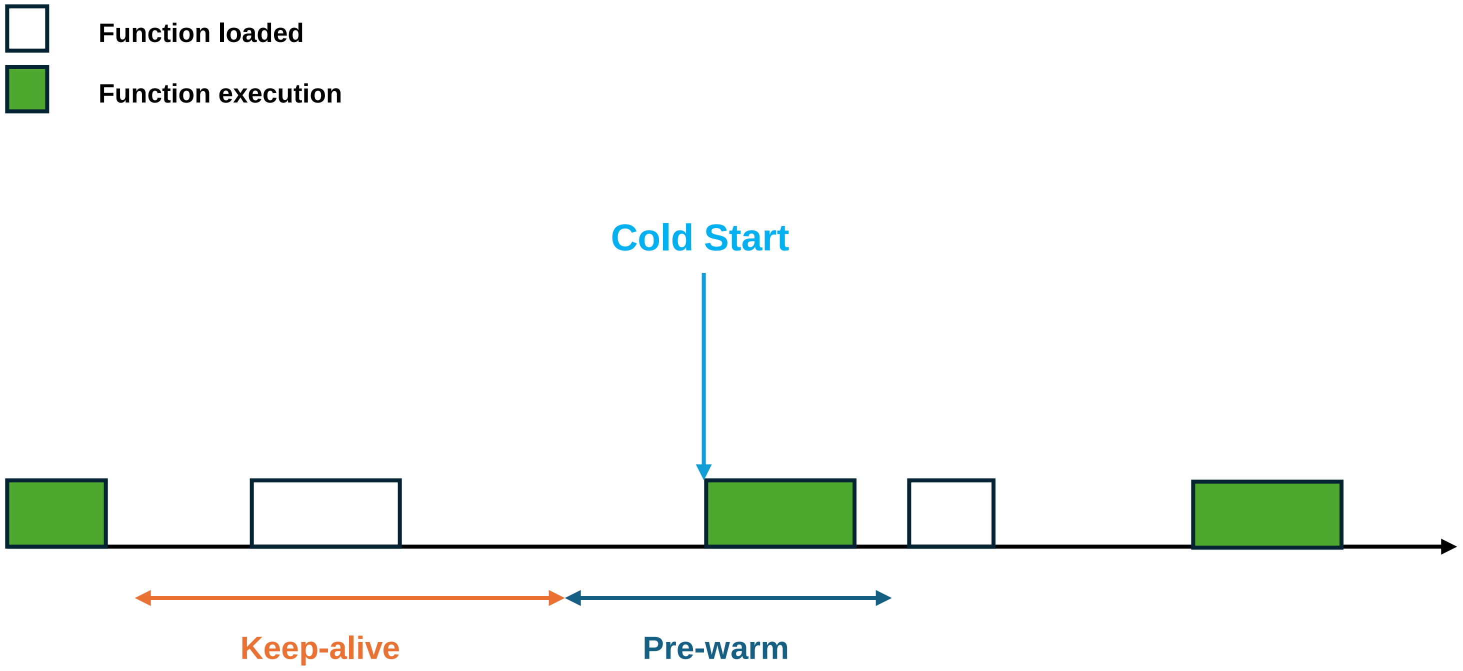
System CPU utilization.

Source: Wang et al., Uncovering The Impact of Bursty Workloads on System Performance in Serverless Computing, ISNCC, 2024

Northeastern University

Challenges

Cold-start latency.



A cold starts, before a pre-warm, and after a keep alive.

Northeastern University

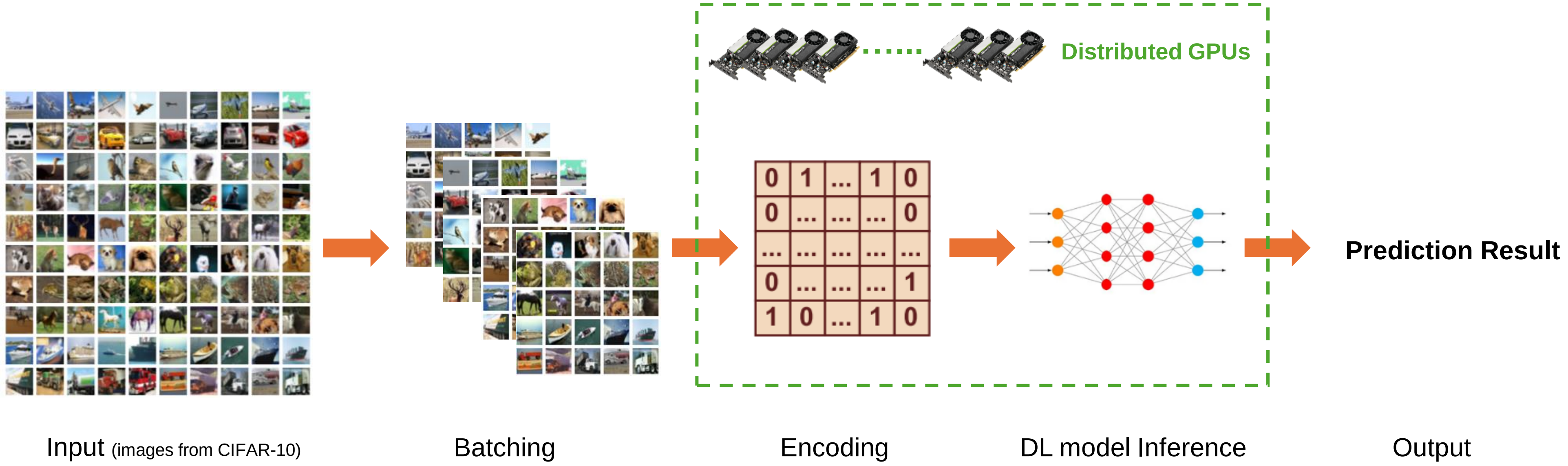
Challenges

Resource over/under-provisioning (CPU, GPU, memory, I/O and network bandwidth, etc.).

Resource under-provisioning	→	**Resource contention**	→	**Performance degradation**
Resource over-provisioning	→	**Resource wastes**	→	**High billing costs**

Challenges

Stateful workflows in distributed AI model inference systems.

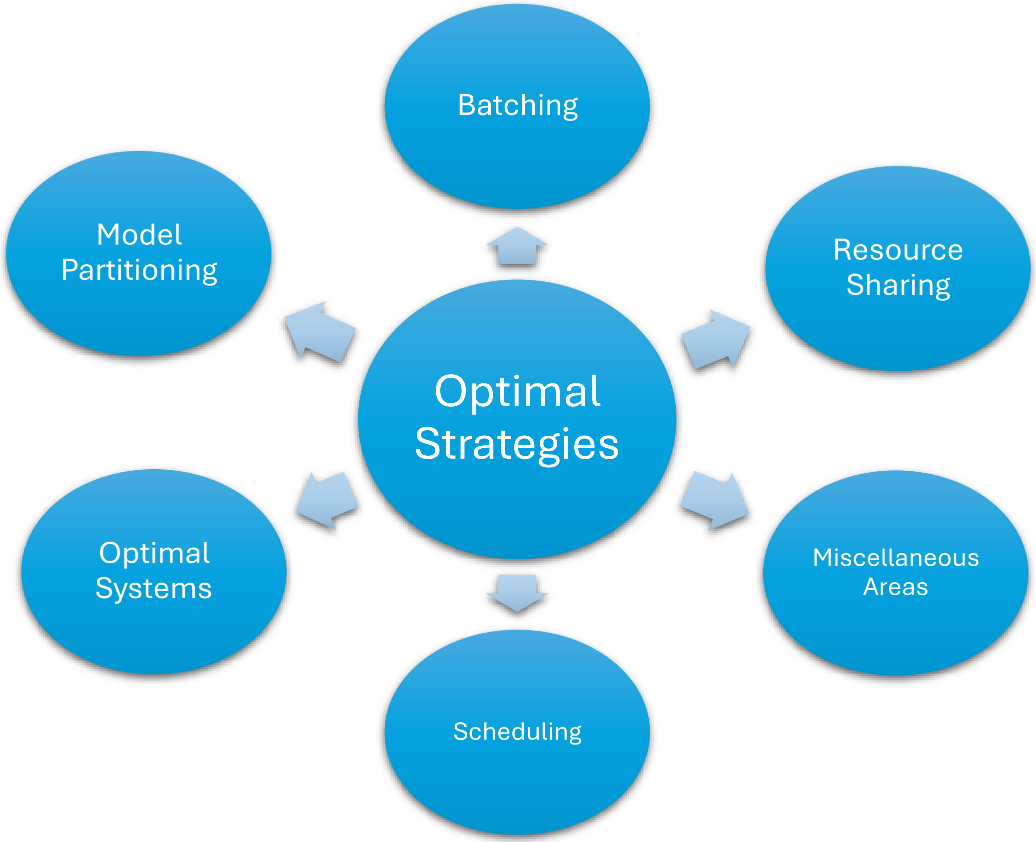


A multi-stage deep learning model serving process.

Northeastern University

Optimal Strategies

Optimal strategies to address the forementioned challenges in scalable AI model inference deployment with serverless paradigm.



Optimal Strategies

Batching and scheduling are commonly used in the context of bursty workloads.

BARISTA	Online resource configurations	Dynamic resource allocation
AYCI	Offer various DL inference configurations	Automate performance evaluation
MArK	Dynamic batching & auto-scaling framework	Recommend small IaaS instances with GPUs
BATCH	Adaptive batching framework	Dynamically calculate the optimal batch size
INFless	Heterogenous hardware configurations & workload prediction	Reduce cold-start & resource wastage
JointBatching	Batching & multi-processing with detection and optimization	Reduce latency under bursty workloads

Optimal Strategies

Model partitioning is used to address the resource demands of large AI models.

MOPAR	Vertically partition the model into slices of analogous layers Data compression & shared memory	Optimize resource usage and reduce latency
Gillis	Partition the model across multi-serverless functions	Ensure optimal latency and SLOs
MLModelComposition	Decompose ML models into slices to execute the large inference tasks as multiple serverless functions	Optimal usage of both storage and memory

Optimal Strategies

Resources sharing (GPUs, memory, networks, and containers).

FaST-GShare	Limits and isolates spatio-temporal resources for GPU multiplexing Allocates executions across GPU nodes	Ensure maximum GPU utilization with SLOs requirements
SMSS	Log-based workflow runtime Two-layer GPU sharing mechanism	Reduce cold-start using inter-/intra-model GPU sharing
Tetris	Dynamic memory-efficient tensor sharing	Performance-cost tradeoff & scalability
GPUColdStarts	Remote memory pooling & hierarchical sourcing through GPU nodes before spawning instances on other nodes	Minimize redundant DL model transformations through download sharing
Fifer	Bin packing, function-aware container scaling, and batching Proactively spawn containers with SLOs requirements	Minimize cold-start latency
Optimus	Inter-function model transformation within container operations	Ensure rapid transitions between models within a container

Optimal Strategies

Optimal resource management systems are designed to recommend optimal configurations.

INFaaS	Generates model variants and creates performance cost profiles across different hardware platforms.	Enable dynamic and efficient selection of optimal variant to meet specific application requirements
AMPS-Inf	Formulates and solves a Mixed-Integer Quadratic Programming (MIQP) problem to partition models and provisions resources	Minimize costs with SLOs requirements for large-scale distributed ML inferences

Optimal Strategies

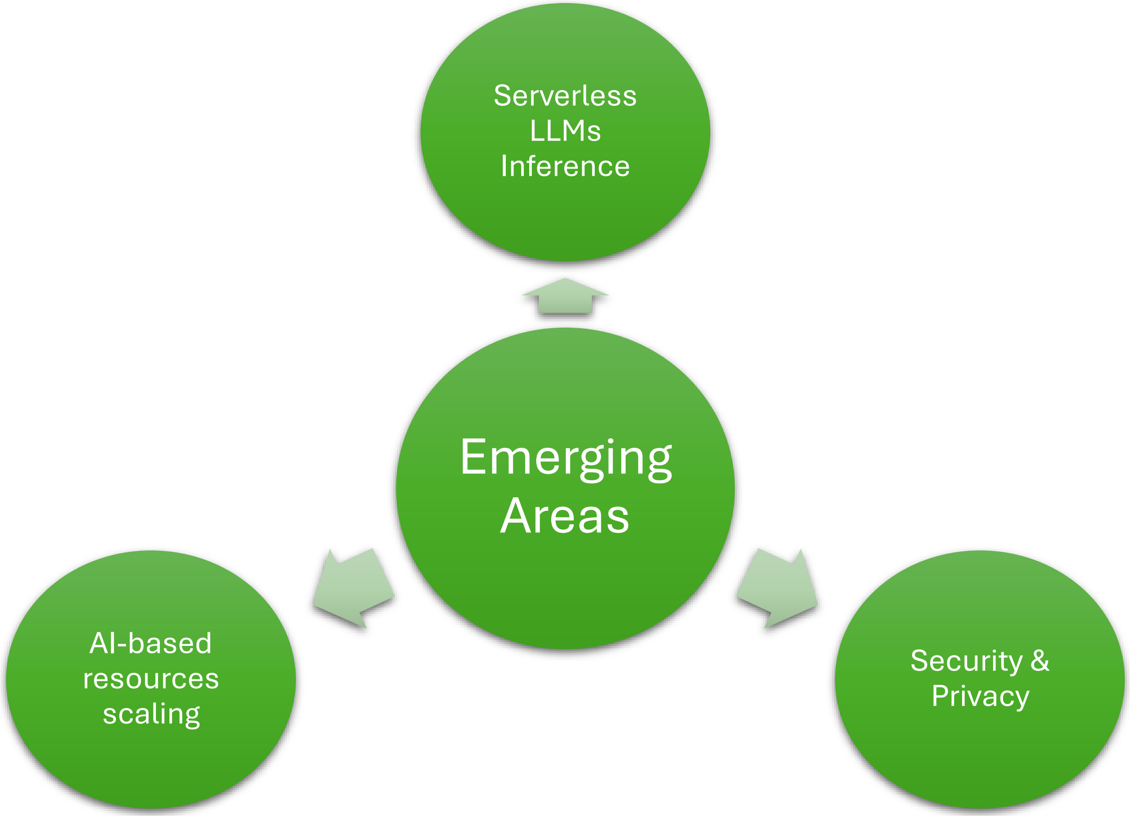
Miscellaneous areas.

●	FSD-Inference	Enable inter-process communication (IPC)	High parallelism with serverless paradigm for ML inference tasks
●	AsyFunc	Separates resource-intensive tasks from higher ones using asymmetric functions and function fusion	Reduce cold-start latency
●	MLFaaS	Generalize ML inference pipelines with AI-based framework Recommend optimal function compositions the pipeline	Minimize the response time of ML inference

Northeastern University

Emerging Research Areas

Serverless LLMs inference.



Northeastern University

Emerging Research Areas

Serverless LLMs inference.

AWS Bedrock	Offers a suite of LLM foundation models	Simplify the complexity of LLM inference and cloud management
Microsoft Azure AI Studio	Pay-as-you-go, token-based billing model Enable users deploy LLM models as serverless APIs	Simplify the complexity of LLM inference and cloud management & save billing costs
LoRAX	A large-scale fine-tuned LLM inference framework using shared GPU resources, continuous batching	A high throughput and low latency system
ServerlessLLM	Checkpointing & multi-tiered model loading	Reduce cold-start latency & speed up model loading
ENOVA	Distributed LLMs inference on multi-GPU nodes	Recommend optimal hardware configurations

Northeastern University

Emerging Research Areas

AI-based scaling.

ServingDI	Hybrid scheduler with deep reinforcement learning techniques	Enable optimal container allocation
Fifer	Function-aware container scaling & LSTM-based request batching	Reduce cold-start latency
Gillis	Encodes partitioning policies into a neural network	Iteratively optimize inference cost and latency

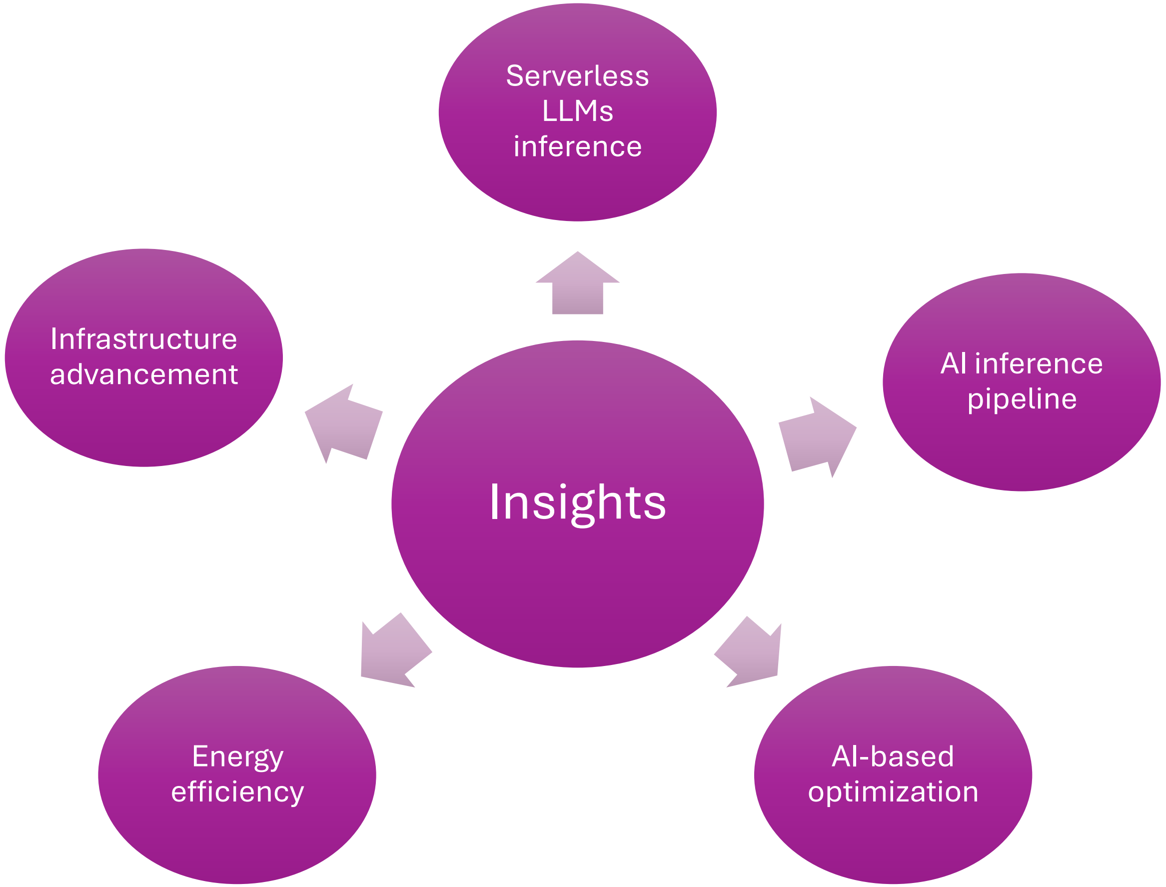
Emerging Research Areas

Security and privacy in edge computing.

●	TrustedLLMInference	Adopts blockchain technology to secure distributed AI inference	Guarantee trust, verifiability, and security for privacy-sensitive tasks in distributed inference.
●	MLEdge	Efficiently deploy ML models with sensitive data being processed locally	Reduce inference latency while preserving privacy

Insights

Serverless LLMs inference.



Insights

Serverless LLMs inference.

- Limitations of existing serverless AI inference frameworks
  - Cold-start latency
  - Resource constraints (memory, storage)
  - Lack of fine-grained control over resources (GPU, memory)
  - Vendor lock-in & high billing costs

- Opportunities
  - Cold-start mitigation
  - Preemptive prediction, scaling, and scheduling for dynamic bursty workloads
  - Distributed multi-GPU management

Insights

Infrastructure advancement.

- Limitations of existing serverless AI inference infrastructures
  - Limitations to perform large-scale inference due to the stateless nature
  - Delays in real-time inference tasks because of continuous access requirements for GPUs

- Opportunities
  - Integration of GPU hardware and AI inference chips
  - Fine-grained control over hardware resources on the cloud

Insights

Energy efficiency.

- Limitations of existing serverless AI inference serving systems
    - Huge energy consumption for large and complex AI model inference (GPU)
    - Energy wastage due to lack of fine-grained control over GPUs
    - Redundant energy wastes because of horizontally scaling in serverless paradigm

- Opportunities
    - Cooling systems for real-time AI inference datacenters
    - Energy-aware scheduling for AI inference models
    - AI model quantization and pruning mechanisms

Insights

AI model inference pipelines.

- Limitations of existing serverless AI inference pipelines
  - Modular, stateless, and event-driven nature of serverless paradigm

- Opportunities
  - Efficient scaling mechanisms
  - Advanced model partitioning strategies
  - AI-based pipeline optimization

Insights

AI-based optimization strategies.

- Limitations of AI-based optimization
  - Resource unpredictability
  - Real-time inference adaptability

- Opportunities
  - Offload AI inference tasks from centralized cloud servers to edge devices
  - Local data processing on edge devices
  - Privacy-sensitive tasks with Federated Learning

Conclusions

- Identify 31 top-tier works among existing literature
- Mark the first comprehensive survey on scalable AI inference with serverless computing
- Analyze challenges & optimal strategies
- Offer valuable insights for both academia and industry

Q & A

Thank you!

Northeastern University

LVX
VERITAS
VIRTVS

Northeastern
University