# Article information:

Serving Heterogeneous Machine Learning Models on Multi-GPU Servers with Spatio-Temporal Sharing | USENIX
<https://www.usenix.org/conference/atc22/presentation/choi-seungbeom>

# Article summary:

1. This paper proposes a new inference scheduling framework for multi-model ML inference servers to address the two requirements of providing a bounded latency and serving multiple heterogeneous ML models.

2. The proposed framework exploits hardware support for spatial partitioning of GPU resources with spatio-temporal sharing, creating a new abstraction layer of configurable GPU resources.

3. The prototype implementation shows that the proposed spatio-temporal scheduling enhances throughput by 61.7% on average compared to the prior temporal scheduler, while satisfying SLOs.

# Article rating:

Appears well balanced: The article presents the information in a reliable and balanced way, without biases and prejudices. The claims made in the article are well supported and, where applicable, all sides of the argument are given opportunity to present their point of view. The article appears trustworthy and reliable.

# Article analysis:

The article is written in an objective manner and provides evidence to support its claims. It presents both sides of the argument equally and does not appear to be biased towards any particular point of view or opinion. The authors provide detailed information about their proposed framework and its advantages over existing solutions, as well as results from their prototype implementation which demonstrate its effectiveness in improving throughput while meeting SLOs.

The article does not appear to contain any unsupported claims or missing points of consideration, nor does it contain any promotional content or partiality towards any particular viewpoint or opinion. All potential risks are noted and discussed in detail, and all counterarguments are explored thoroughly.

In conclusion, this article appears to be trustworthy and reliable, providing an unbiased overview of the proposed framework with evidence to support its claims.

# Topics for further research:

* Network throughput optimization
* Service Level Agreement (SLA) compliance
* Network latency reduction
* Network traffic management
* Network resource allocation
* Network performance monitoring

# Report location:

<https://www.fullpicture.app/item/51c220e8ebbe2529b982b3b58f8c2c88>