# A Method For Scheduling Multi-Model AI Workloads Onto Multi-Chiplet Modules

Tech ID: 33870 / UC Case 2024-978-0

## BRIEF DESCRIPTION

This technology introduces an advanced scheduling strategy for optimizing multi-model AI workloads on heterogeneous chiplet-based multi-chip modules (MCMs), aiming at maximizing performance efficiency.

## FULL DESCRIPTION

UCI Researchers have developed technology addressing the challenge of efficiently scheduling multi-model AI workloads on heterogeneous chiplet-based MCMs. It proposes a bi-level optimization problem that includes time partitioning for reconfiguration of MCM chiplets and spatial mapping of sub-model workloads to chiplets. The solution aims to enhance in-package data reuse, reduce off-chip traffic, and improve overall performance efficiency in terms of energy efficiency and latency.

## SUGGESTED USES

- AI hardware for edge to cloud computing, enhancing compute capability.
- AI accelerators for large language models and multi-model deployments such as AR/VR.
- Energy and latency-efficient AI inference engines for scalable multi-chip architectures.
- Optimization software for AI workload deployment on heterogeneous computing platforms.

## ADVANTAGES

- Addresses workload heterogeneity in multi-model AI workloads with a heterogeneous chiplet-based approach.
- Enhances in-package data reuse and reduces off-chip traffic through inter-layer pipelining.
- Employs advanced scheduling techniques including dynamic chiplet regrouping and resource allocation trees.
- Significantly reduces energy-delay product (EDP) and latency compared to homogeneous MCMs.
- Future-proofs for emerging AI workloads with an extendable and scalable solution.

## PATENT STATUS

Patent Pending

## STATE OF DEVELOPMENT

## CONTACT

Edward Hsieh
hsiehe5@uci.edu
tel: 949-824-8428.



## OTHER INFORMATION

## CATEGORIZED AS

- **Communications**
  - Networking
- **Computer**
  - Other
- **Semiconductors**
  - Other

## RELATED CASES

2024-978-0

Validated in laboratory environment

UCI Beall
Applied Innovation

5270 California Avenue / Irvine,CA
92697-7700 / Tel: 949.824.2683