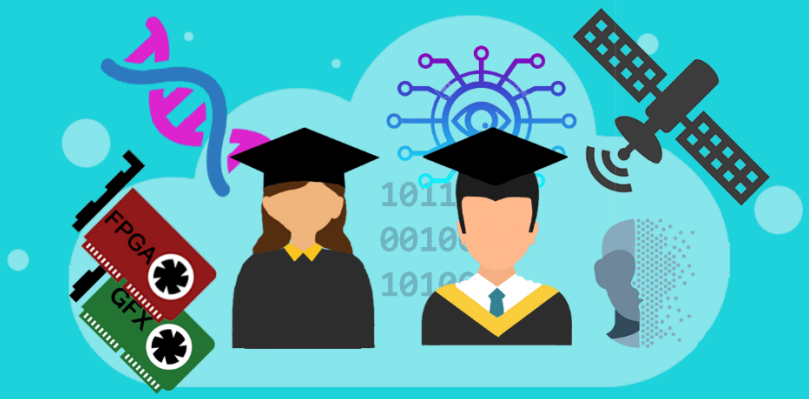




Architectural Exploration of 3D-Integrated GPGPUs

Motivation

Modern GPUs achieve high computational throughput by integrating a large number of parallel processing units with increasingly complex memory hierarchies and interconnection networks. As these architectures scale, **data movement, memory bandwidth, communication latency, and physical interconnect costs** increasingly limit both performance and energy efficiency.

3D integration provides the opportunity to distribute compute, memory, and communication resources across multiple vertically interconnected silicon tiers. The potential of 3D integration extends beyond reducing chip footprint: dense vertical connectivity can fundamentally change communication costs between architectural components, enabling organizations that would be impractical in a conventional 2D implementation [1],[2].

Multiple Levels of Memory – Logic - I/O

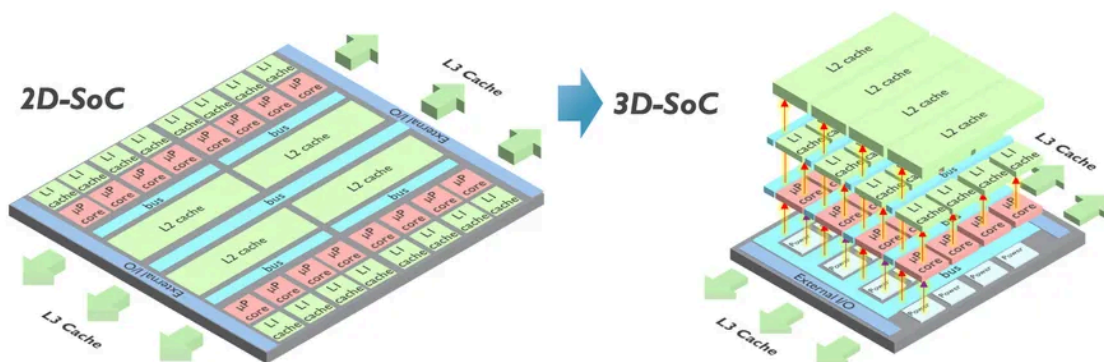


Figure 1: Transition from traditional 2D architectures to 3D [1]

The important question is therefore not simply *how much faster an existing GPU would become if it were stacked in 3D*, but rather:

How should the GPU architecture change to exploit the opportunities provided by 3D integration?

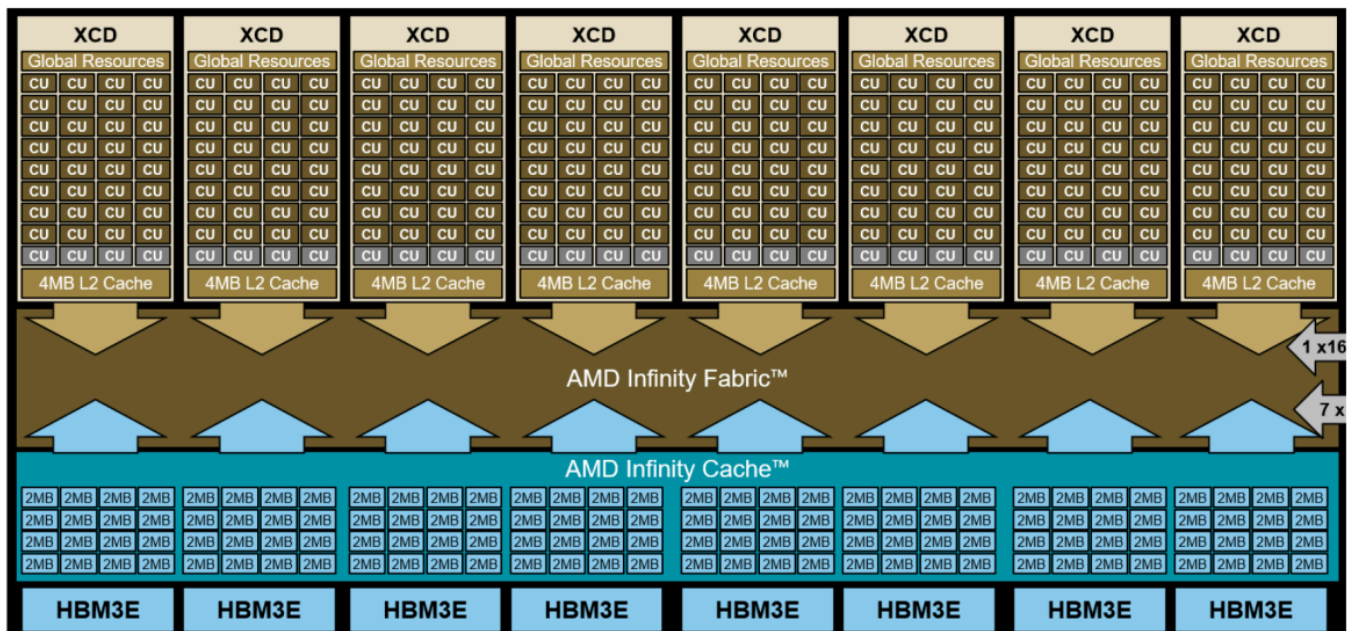


Figure 2: Conceptual block diagram of an AMD Instinct™ MI350 Series GPU

Scope

The objective of this diploma thesis is to investigate the **architectural opportunities introduced by 3D integration in GPGPU systems**, with particular emphasis on the interaction between computation, memory hierarchy, and on-chip communication.

The first part of the thesis will focus on the **characterization of conventional 2D GPGPU architectures**. The student will study representative GPU workloads and identify the major sources of performance and energy overhead. Particular attention will be given to determining where data movement occurs, which architectural components exchange the largest volumes of data, and which parts of the memory and communication hierarchy constitute bottlenecks.

This analysis will investigate questions such as whether execution is limited by computational throughput or by the ability to supply data to the processing units, how frequently different levels of the memory hierarchy are accessed, how much traffic is generated between GPU cores and shared caches, and how these characteristics vary across different classes of workloads.

The study will then examine the **physical implications of these communication patterns**. Frequently communicating components may be logically close in the architecture while remaining physically separated in a conventional 2D implementation. The thesis will therefore investigate the contribution of interconnect distance, memory organization, cache placement, and network topology to overall latency and energy consumption.

Opportunities enabled by 3D integration

Based on the identified 2D bottlenecks, the thesis will investigate which of them can be alleviated through **vertical integration**. An initial analysis could consider the direct effects of distributing an otherwise similar GPU architecture across multiple tiers. This will establish the potential benefits of shorter communication paths, increased connection density, and additional memory bandwidth independently of more significant architectural modifications.

The main focus, however, will be on identifying **architectural opportunities that become possible specifically because of 3D integration**. Examples include placing memory resources physically closer to the processing units that use them [2], providing substantially greater bandwidth between compute and memory, reorganizing the cache hierarchy, increasing memory capacity without proportionally increasing lateral communication distances, and reconsidering the organization of the GPU interconnection network.

An important aspect of the thesis will therefore be the exploration of the **compute–memory relationship**. Different workloads place significantly different demands on caches, shared memories, and off-chip memory. The student will investigate whether 3D integration can enable memory organizations that adapt more effectively to these requirements and whether conventional assumptions regarding cache hierarchy, banking, sharing, and placement remain appropriate. Similarly, the thesis will examine how 3D integration affects **GPU communication and scalability**. As the number of processing units increases, conventional on-chip networks may become increasingly expensive in terms of latency, energy, and physical distance. Vertical connectivity introduces an additional communication dimension, potentially enabling new organizations of processing units, caches, and network resources.

Architectural Trade-offs

The opportunities provided by 3D integration must also be considered together with the constraints it introduces. The relationship between communication efficiency and power density is particularly important. Placing highly active compute and memory structures close together may reduce data-movement energy while simultaneously creating thermal hotspots and increasing thermal coupling between tiers.

The thesis will therefore investigate the trade-offs between:

- communication distance, bandwidth, and energy;
- compute and memory placement
- cache capacity, organization, and accessibility;
- performance improvements and power density;
- number and organization of tiers;
- homogeneous and heterogeneous tier organizations;
- architectural scalability and thermal constraints.

RELATED WORKS:

- [1] [P. J. NAIR ET AL., "EARLY SILICON OF RAPTOR: THE FIRST 3D-DRAM ACCELERATOR FOR GENERATIVE INFERENCE," 2026 INTERNATIONAL SYMPOSIUM ON COMPUTER ARCHITECTURE \(ISCA\)](#)
- [2] [G. H. LOH, "3D-STACKED MEMORY ARCHITECTURES FOR MULTI-CORE PROCESSORS," 2008 INTERNATIONAL SYMPOSIUM ON COMPUTER ARCHITECTURE](#)
- [2] [C. T. DO, Y. -H. GONG, C. H. KIM, S. W. KIM AND S. W. CHUNG, "EXPLORING THE RELATION BETWEEN MONOLITHIC 3D L1 GPU CACHE CAPACITY AND WARP SCHEDULING EFFICIENCY," 2019 IEEE/ACM INTERNATIONAL SYMPOSIUM ON LOW POWER ELECTRONICS AND DESIGN \(ISLPED\)](#)

CONTACT INFORMATION:

- Alexis Maras, Ph.D Student, NTUA: (amaras@microlab.ntua.gr)
- Panagiotis Eleftherakis, Ph.D. Student, NTUA: (pelef@microlab.ntua.gr)
- Maria Zerva, Research Engineer, Barcelona Supercomputer Cluster: (maria.zerva@bsc.es)
- Associate Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)