Investigating Memory Hierarchy In The PlayStation 5

Introduction
The Sony PlayStation 5 represents a highly sophisticated system that exemplifies advanced memory hierarchy design in modern gaming consoles, engineered to provide exceptional performance while adhering to stringent thermal and power constraints. This system is particularly relevant to the field of computer architecture, as it showcases an innovative approach to data management characterized by the integration of an ultra-high-speed solid-state drive (SSD) directly into its core memory architecture. The focus of this analysis is to examine how the hierarchy of caches, unified GDDR6 memory, and custom SSDs is structured to leverage principles of locality and reduce latency, thereby facilitating new paradigms in immersive game design.

Background / Related Works
The design of memory hierarchies is fundamentally influenced by the principles of temporal and spatial locality. Temporal locality is effectively managed through multi-level CPU caches (L1, L2, L3), which are constructed from fast Static RAM (SRAM) and are designed to store recently accessed data. Spatial locality is typically addressed by employing cache prefetching techniques for adjacent memory blocks, a practice that has now been further developed to include storage-level prefetching in contemporary systems. The trade-offs among various memory technologies are significant: while SRAM is characterized by its speed, it comes at a higher cost and greater power consumption; DRAM provides a desirable balance between capacity and bandwidth; and NAND flash delivers extensive, non-volatile storage at a lower cost, albeit with increased latency (Harvie, 2023).

The PlayStation 5 features a custom AMD System-on-a-Chip (SoC) that integrates Zen 2 CPU cores and RDNA 2 GPU cores, representing a contemporary application of advanced computing principles (Leadbetter, 2020). The architecture of the PS5's CPU and GPU is organized as follows: the L1 Cache comprises 32KB instruction and 32KB data per core, structured as 8-way associative; the L2 Cache provides 512KB per core, also 8-way associative; and there is a Unified L3 Cache of 8MB shared between the CPU and GPU (WikiChip, 2025). Additionally, the GPU architecture includes a L0 Cache of 16KB-32KB per compute unit, an L1 Cache of 128KB per shader engine, and an L2 Cache of 4MB shared across the GPU (Chester, 2021). The system is equipped with 16GB of RAM and an SSD. Notably, the architectural design, as elucidated by lead architect Mark Cerny and examined by various technical experts, transcends conventional hierarchies by conceptualizing the SSD as a direct, high-throughput extension of the main memory, rather than merely as storage (Strickland, 2020). Recent investigations into memory hierarchy optimization emphasize system-DRAM co-design approaches that facilitate enhanced cooperation between memory controllers, processors, and storage systems. Academic research underscores the significance of reducing refresh overhead, augmenting parallelism, and minimizing data movement across memory tiers (Ge et al., 2025). The gaming industry has increasingly embraced unified memory architectures and custom SSD solutions to address traditional loading bottlenecks effectively (Strickland, 2020).

Methodology / System Design
Based on the analysis of the PS5's existing Unified Memory Architecture (UMA), this report proposes a strategic redesign: the implementation of a Tiered Unified Memory Architecture, which was used in the XBOX series X (Tuttle, 2020). This design seeks to retain the core benefits of a unified pool while mitigating the potential for bandwidth contention between the CPU and GPU by creating dedicated bandwidth tiers within the same physical memory (Reddit, 2020).

Some assumptions –

1. The overall memory capacity remains 16 GB, and the SSD still works like an extension of the main memory.
2. The custom AMD SoC and I/O complex remain largely unchanged, requiring only a modified memory controller.
3. The target cost and power envelope for the console remain the same as the original design.
4. Game developers would be provided with a refined software development kit (SDK) to assign memory priorities easily.

The proposed design modifies the memory controller to partition the unified GDDR6 pool into two logical tiers with different performance characteristics, while maintaining a unified address space. The following table shows the new architecture over the original one:

Features	Original PS5 Design	Proposed Tiered Design
Total Capacity (main memory)	16 GB Unified GDDR6	16 GB Unified GDDR6
Memory Bus	256-bit	320-bit
Tier1 (Fast)	N/A	12 GB at 504 GB/s
Tier2 (Standard)	16 GB at 448 GB/s	4 GB at 144 GB/s
Primary Use Case	All CPU and GPU data are using only one tier.	Mainly handle the GPU critical works on tier1, along with the CPU tasks using tier2.

The tiered UMA offers improvements in the following things-

1. **Performance Improvement**: One of the primary advantages of the proposed tiered model is the enhancement in performance. In the original UMA design, CPU-intensive tasks could inadvertently consume memory bandwidth, potentially leading to increased latency for the GPU. The tiered model ensures that the GPU has guaranteed access to a significant portion of the total system bandwidth (504 GB/s) for its most critical operations, effectively isolating it from CPU activities. This results in more consistent and predictable frame times, which are essential for delivering smooth gaming experiences.
2. **System Cost Improvement**: The proposed design also presents an opportunity for improved cost efficiency. The Standard Tier (Tier 2) has lower bandwidth requirements, which could enable the use of a reduced number of cost-effective, lower-clocked GDDR6 chips for this segment of the memory pool. Alternatively, it may permit the implementation of different memory solutions for CPU-centric tasks. The design maintains the use of the same GDDR6 chips while regulating bandwidth allocation across the memory pool, which represents a cost-neutral adjustment from a hardware standpoint, yet facilitates a considerable performance enhancement.
3. **Energy Use**: The energy consumption associated with this design would remain largely consistent. Nevertheless, by segregating memory traffic, the memory controller could potentially implement more effective power-gating strategies for specific portions of the memory bus. Directing the majority of CPU traffic to a smaller, lower-bandwidth pool could diminish the energy expenditures associated with memory accesses for standard computing tasks, resulting in a slight overall enhancement in power efficiency.

Discussion

The proposed Tiered UMA architecture is poised to significantly enhance the performance profile of the PlayStation 5. By ensuring the GPU's access to a high-bandwidth pool of 12 GB at 504 GB/s, this design directly addresses a critical limitation associated with traditional UMA systems: bandwidth contention. This enhancement is expected to yield more consistent performance, improve stable frame rates in complex scenes, and reduce instances of micro-stuttering, as the GPU's throughput becomes insulated from CPU activity (Reddit, 2020). The prioritization logic, implemented via an enhanced memory controller and software development kit (SDK), will empower developers with precise control over performance, allowing them to ensure that essential rendering assets consistently reside in the high-speed tier. This approach elevates the system's functionality from merely providing raw bandwidth to delivering intelligent and predictable bandwidth management.

Nevertheless, this proposal introduces specific trade-offs and limitations. The primary consideration is an increase in complexity, a pivotal aspect that the original UMA sought to minimize. Developers will assume responsibility for memory tiering, which is not a trivial task. It necessitates careful profiling and asset tagging to prevent the misallocation of latency-sensitive data into the slower tier, which could inadvertently impair performance. Additionally, the fixed sizes of the memory tiers—12 GB for the high-speed tier and 4 GB for the standard tier—present inherent inflexibility. A game with exceptionally high CPU memory requirements, such as large-scale strategy games, may encounter constraints due to the limitations of the 4 GB standard tier, while simpler games may not fully utilize the faster tier’s capacity. As there are some games that need CPU memory more than the GPU, in that case, the SSD can back up, as we assume that it will work like an extension of the main memory, but this must be designed by the developers. A novel concept examined within this framework is the "unified-but-partitioned" address space. Unlike discrete architectures, the tiers remain part of a single coherent pool, allowing the CPU to access Tier 1 and the GPU to access Tier 2 when necessary. This maintains the inherent advantages of a UMA while introducing an added layer of performance governance.

While the continual innovation embodied in architectures like the Tiered UMA drives enhanced performance, it also raises significant concerns regarding sustainability. The production of increasingly sophisticated and specialized semiconductors, such as advanced memory controllers and heterogeneous memory pools required for this design, leads to substantial consumption of energy, water, and rare earth elements, contributing to a considerable embedded carbon footprint even before the device is deployed. Furthermore, a business model that promotes frequent generational upgrades exacerbates the global issue of electronic waste. Although these technological advances can enhance energy efficiency during use, this benefit is frequently countered by the overall elevation in performance, known as the "rebound effect," alongside the environmental costs associated with production. Thus, the industry confronts a critical challenge: striving for peak performance to deliver immersive experiences must be balanced with a commitment to fostering a sustainable lifecycle for increasingly essential technology.

Conclusion

The insights gained from this analysis indicate that a pure Unified Memory Architecture (UMA), while inherently elegant and efficient, presents a potential trade-off between developer simplicity and assured performance. This challenge can be effectively addressed through the introduction of intelligent tiering within the unified memory pool. The proposed Tiered UMA approach offers significant benefits to system design by ensuring more predictable and high-priority bandwidth allocation for the GPU, ultimately enhancing the user experience through smoother and more consistent gameplay. Looking ahead, it would be prudent to investigate the development of a hardware- and software-controlled adaptive tiering system. Such a system would dynamically adjust the size of each memory pool in response to real-time workload demands, moving beyond fixed allocations. This approach aims to strike an optimal balance between flexibility and performance isolation, thereby further improving system efficiency and user satisfaction.

References

Chester, A. (2021, April 6). *RDNA 2 deep dive*. CustomPC. https://www.custompc.com/rdna-2-unravelled

Ge, Z., Lim, H. B., & Wong, W. F. (2025, January 1). Memory Hierarchy Hardware-Software Co-design in Embedded Systems. *The Open University*, *1*, 9. https://core.ac.uk/download/pdf/4384366.pdf

Harvie, L. (2023, August 9). *Embedded Systems Memory Types: Flash vs SRAM vs EEPROM*. Medium. Retrieved August 28, 2025, from https://medium.com/@lanceharvieruntime/embedded-systems-memory-types-flash-vs-sram-vs-eeprom-93d0eed09086

Leadbetter, R. (2020, March 29). *Inside PlayStation 5: the specs and the tech that deliver Sony's next-gen vision*. Eurogamer. Retrieved August 28, 2025, from https://www.eurogamer.net/digitalfoundry-2020-playstation-5-specs-and-tech-that-deliver-sonys-next-gen-vision

Reddit. (2020, March 17). *Why Xbox Series X's Dumb 10+6GB Memory Configuration Isn't As Dumb As You Think: r/hardware*. Reddit. Retrieved August 29, 2025, from https://www.reddit.com/r/hardware/comments/fjy1sw/why_xbox_series_xs_dumb_106gb_memory/

Strickland, D. (2020, March 18). *Understanding the PS5's SSD: A deep dive into next-gen storage tech*. TweakTown. Retrieved August 28, 2025, from https://www.tweaktown.com/news/71340/understanding-the-ps5s-ssd-deep-dive-into-next-gen-storage-tech/index.html

Tuttle, W. (2020, March 16). *Xbox Series X: A Closer Look at the Technology Powering the Next Generation*. Xbox Wire. Retrieved August 29, 2025, from https://news.xbox.com/en-us/2020/03/16/xbox-series-x-tech/

WikiChip. (2025, April 27). *Zen 2 - Microarchitectures - AMD*. WikiChip. Retrieved August 29, 2025, from https://en.wikichip.org/wiki/amd/microarchitectures/zen_2