Abstract

Optimizing Memory Management for Disaggregated Architectures

Yupeng Tang

2024

The increasing demand for scalable and efficient data center architectures has led to the adoption of resource disaggregation, which separates compute, memory, and storage resources across various interconnects. This paradigm shift from traditional monolithic server architectures allows for more flexible resource allocation and utilization. Memory disaggregation, in particular, addresses the bottleneck issues of traditional setups by decoupling memory resources, presenting them as pooled resources accessible on demand. This approach enhances efficiency, scalability, and adaptability, especially for memory-intensive workloads.

However, transitioning existing applications to a disaggregated architecture presents significant challenges due to the mismatch between current cloud stacks designed for monolithic systems and the requirements of disaggregated systems. These challenges span across different layers of the stack, including application interfaces, OS support, performance overheads, and the limitations of existing interconnect technologies. This dissertation focuses on addressing these challenges, particularly in the context of memory management within disaggregated architectures.

Our approach involves a comprehensive examination of the requirements for successful disaggregation, proposing strategies to mitigate performance penalties and enhance resource management. By adopting a top-down perspective, we aim to bridge the gap between service layers and core hardware elements, ultimately facilitating the transition to disaggregated data center architectures.

Optimizing Memory Management for Disaggregated Architectures

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Yupeng Tang

Dissertation Director: Anurag Khandelwal

Dec, 2024

Contents

List of Figures

List of Tables

Acknowledgements

A lot of people are awesome. Probably your family, friends, advisor, and that one super special high school teacher who believed in you.

Chapter 1

Introduction

The growing demand for scalable and efficient data center architectures has led to the emergence of resource disaggregation [1–9]. This modern paradigm represents a significant shift from traditional monolithic server architectures. In conventional setups, servers are typically equipped with a fixed combination of compute, memory, and storage resources. In contrast, resource-disaggregated systems physically separate these resources and distribute them across various interconnects, such as networks [1–3], CXL [10, 11], and others. This separation allows for more flexible resource allocation and utilization.

Within the broader context of resource disaggregation in modern data center architectures, **memory disaggregation** [4–9] plays a crucial and foundational role. In traditional monolithic architecture, memory often becomes a bottleneck, limiting the scalability and adaptability of applications. This issue has been frequently observed and reported in production data centers [12–21]. By decoupling memory resources from compute and storage elements and presenting them as pooled, disaggregated resources [22, 23], data centers can achieve increased efficiency, scalability, and adaptability. Memory-intensive applications [24–26] can access the memory they need on demand, without being constrained by the limitations of individual servers. Memory disaggregation is the first step toward realizing the full potential of resource disaggregation, enabling data centers to efficiently allocate and utilize resources based on dynamic application needs.

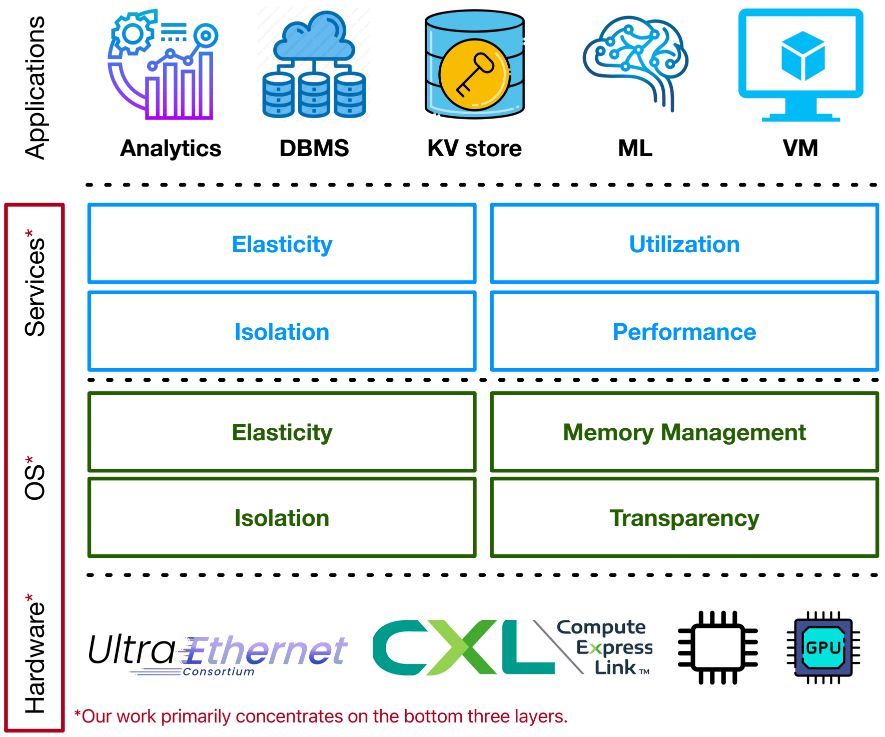


Fig. 1.1: **Cloud Stack of Disaggregated Architecture.**

1.1 Limitations of Existing Approaches

While memory disaggregation offers numerous advantages, transitioning existing applications to this architecture is far from straightforward. Recent research has explored various approaches to tackle this challenge. Some efforts focus on optimizing applications for disaggregated memory [27–30], while others aim to transparently port applications, offloading the responsibility for mitigating performance penalties—stemming from mismatches between disaggregated architectures and traditional software interfaces—to the service or operating system layer [1, 2, 31–34]. Meanwhile, hardware innovations further complicate the situation. Different interconnects, such as Ethernet and CXL, offer diverse interfaces and performance characteristics, making it difficult to standardize software stack designs.

The core issue lies in the fundamental mismatch between the existing cloud stack, designed for monolithic architectures, and the requirements of disaggregated architectures (Figure 1.1). The current cloud and hardware stacks are not inherently aware of the unique characteristics of disaggregated memory, leading to distinct challenges across different layers of the stack:

Application interface. In disaggregated architectures, applications encounter unique challenges compared to traditional monolithic systems. The key difference lies in resource distribution: compute, memory, and storage are spread across multiple resource nodes rather than centralized within a single server. This distribution necessitates complex communication and data management strategies to address increased latency and resource coordination. Transparency also becomes a critical concern. There is an inherent trade-off between the benefits of disaggregating resources and the associated performance overhead. Application developers may either significantly redesign their applications for more efficient resource utilization and management or rely on disaggregated cloud providers to integrate resource management at the service or operating system layer to ensure a smoother transition.

Operating system support. Unlike monolithic servers where the operating system(OS) manages resources within a single server, the placement and function of the OS in disaggregated architectures are still subjects of debate in both industry and academia. Options include centralizing the OS at a single point [1, 35] in the architecture or disaggregating its functions across different resource nodes [2].

Performance overheads. Transitioning existing applications to a disaggregated architecture transparently introduces a spectrum of performance challenges. These include, but are not limited to, managing memory partitioning [36] and addressing applications with irregular memory access patterns [35]. Various other issues, such as latency sensitivity, bandwidth limitations, and the overhead of remote resource management, compound this complexity. These factors contribute to the overall performance penalty that disaggregated systems must carefully consider and mitigate.

Future interconnects. The use of networks as interconnects for resource disaggregation has been a focus of research in both academia and industry. However, networks face inherent challenges, such as performance slowdowns compared to intra-server resource access and the absence of built-in coherency. Emerging hardware technologies like Compute Express Link (CXL) [10, 11, 37] offer promising improvements, including faster access times and hardware-supported cache coherence. Despite this potential, current hardware prototypes and software support for these technologies are still in early stages, and their full impact on software design remains uncertain.

1.2 Thesis Overview

In this dissertation, we attempt to take a top-down approach and explore the optimal memory management solutions for three most significant layers, i.e. Service, OS and Hardware layers of disaggregated memory architectures.

1.2.1 Service Layer: Memory Management as a Service

As the layer closest to the application, we first explore the design requirements and challenges of providing disaggregated memory management as a service. This service manages a pool of memory resources and exposes them to applications. We propose an end-to-end system design called Jiffy, which enables multiple applications or tasks to efficiently share memory resources in an elastic manner. Jiffy also offers interfaces for several popular data structures, making it easily applicable to existing cloud applications.

1.2.2 OS Layer: In-Network Memory Management

While applications may use Jiffy to manage memory resources, we take a deeper approach by considering the Operating System (OS) as the manager of memory. This allows applications to run transparently while benefiting from the disaggregated architecture. In disaggregated systems, compute and memory resources are decoupled, meaning there is no single host, as in monolithic architectures, to handle the critical task of resource management—traditionally done by the OS. We propose a new generation OS design that embeds OS functionality within the interconnects. Starting with a system called MIND, we address fundamental challenges in memory management, such as memory address translation, memory protection, and cache coherence across multiple hosts. This resource decoupling and in-network memory management perform well for cache-friendly workloads but struggle with cache-unfriendly workloads due to the overhead of interconnect communication. To address this, we developed PULSE, a near-memory accelerator designed from the ground up. PULSE analyzes common pointer traversal applications and identifies a simple, yet effective interface that can be easily integrated into existing cloud applications.

1.2.3 Hardware Layer: Memory Management for Next-Gen Interconnects

In prior work [1, 2], Ethernet has been the most commonly used interconnect for disaggregated data centers. However, with the emergence of new memory interconnects like Compute Express Link (CXL), memory management must be adapted to accommodate these new interfaces. In the context of disaggregated architectures, new challenges arise, such as how applications can effectively utilize multiple tiers of memory. To address this, we begin with a performance evaluation of CXL 1.1 extended memory in a single-host environment and explore how modern data center applications can benefit from such disaggregated memory systems.

1.3 Outline and Previously Published Work

This dissertation is organized as follows. Chapter 2 introduces Jiffy, a distributed memory management system that decouples memory capacity and lifetime from compute in the serverless paradigm. Chapter 3 describes two innovated system design: (1) MIND, a rack-scale memory disaggregation system that uses programmable switches to embed memory management logic in the network fabric. (2) PULSE, a framework centered on enhancing in-network optimizations for irregular memory accesses within disaggregated data centers. Chapter 4 presents our exploration in latest Compute Express Link(CXL) hardware. We conclude with our contributions and possible future work directions in Chapter 5.

Chapter 2 revises material from [36][1]. Chapter 3 revises material from [1][2] and [35][3]. Finally, Chapter 4 revises material from [38][4].

1. Work done in collaboration with Rachit Agarwal, Aditya Akella, and Ion Stoica

2. Work done in collaboration with Seung-seob Lee, Yanpeng Yu, Lin Zhong and Abhishek Bhattacharjee

3. Work done in collaboration with Seung-seob Lee and Abhishek Bhattacharjee

4. Work done in collaboration with the Bytedance Infrastructure team

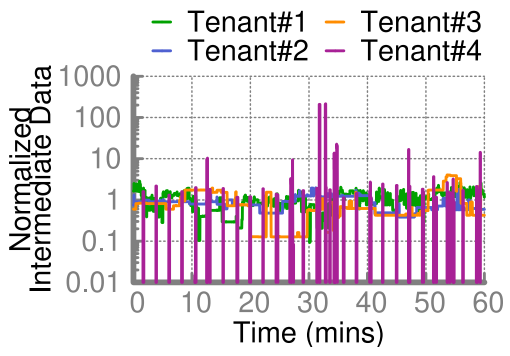
Chapter 2

Service Layer: Memory Management as a Service

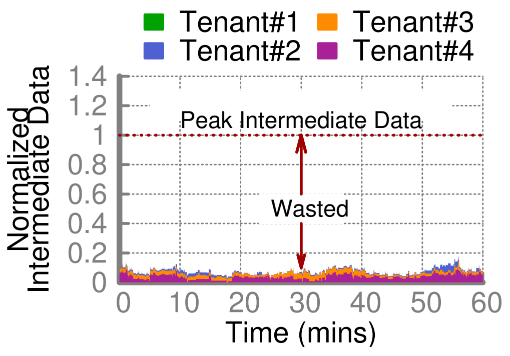
We begin by focusing on the service layer, which sits above the OS layer and plays a critical role in enabling efficient memory sharing across multiple compute and memory nodes in disaggregated architectures. This layer offers greater flexibility than the OS, allowing it to provide adaptable services that cater to the specific needs of different applications. However, this flexibility comes at the cost of potential significant modifications to applications, especially when decoupling storage and compute resources is not straightforward. Without proper decoupling, developers may face substantial challenges in adapting their applications to fully utilize memory management services.

To address this, we start with data analytics applications in serverless computing [39–54], a widely adopted workload in modern data centers. Serverless architectures inherently offer on-demand elasticity, decoupling compute and storage resources logically. Recent advances in serverless analytics have demonstrated the benefits of using serverless architectures for resource- and cost-efficient data analytics. In these systems, remote, low-latency, high-throughput disaggregated memory is used to store intermediate states for inter-task[1] communication and multi-stage jobs, extending the lifetime of data beyond the task that produced it. This natural separation of compute and memory makes serverless computing an ideal candidate for leveraging disaggregated memory

1. Existing distributed programming frameworks, while different in underlying programming models and semantics, share a common structure (Figure 2.2) — the job is split into multiple *tasks*, possibly organized along multiple stages or a directed acyclic graph. Each task generates *intermediate* data during its execution; upon completion, each task partitions its intermediate data and exchanges it with tasks in the next stage.



(a) Intermediate data (normalized by mean usage)



(b) Cumulative intermediate data (normalized by peak usage)

Fig. 2.1: Analysis of production workloads from Snowflake [39] for four tenants over a 1 hour window: (a) the ratio of peak to average storage usage for a job can vary by an order of magnitude during its execution; and (b) provisioning for peak usage results in average utilization $< 10\%$. Across all tenants, the average utilization is 19%.

architectures.

As discussed in [40,42,45,55], these applications handle user requests in the form of jobs, each defining its memory needs upon creation. The dilemma of balancing performance with resource efficiency for job-level memory allocation has been extensively studied [56,57]. If a job is based on average demand, performance may decline during peak demand periods due to inadequate memory, causing data spillage to slower secondary storage, such as SSDs. Conversely, allocating memory for peak demands leads to underutilization of resources when the actual demand is below peak. Evaluations on Snowflake's workload, as shown in [56], indicate a significant fluctuation in the ratio of peak to average demands, sometimes varying by two orders of magnitude within minutes.

Designing a memory management service for such systems is a non-trivial task. We begin by outlining the essential requirements for memory management in disaggregated environments, focusing on the unique challenges posed by disaggregation. We then discuss our efforts to address these challenges and suggest potential future directions for research in this rapidly evolving field.

Elasticity. Memory usage in modern computing is highly variable, with applications facing fluctuating demands [36]. Elasticity enables dynamic memory allocation based on current needs, optimizing resource utilization. Applications like data analytics consist of jobs with multiple tasks that communicate via intermediate memory. Traditional solutions allocate memory at the job level, where jobs specify their requirements before execution, and the system reserves that amount for the job's duration [42]. This approach creates a tradeoff: allocating for average demand risks performance degradation due to swapping data to slower storage (e.g., S3), as shown in Figure 2.1(a), while allocating for peak demand leads to resource waste (Figure 2.1(b)). Recent studies report that intermediate data sizes can vary by orders of magnitude during a job's lifetime [39]. For example, Figure 2.1 shows that in a Snowflake dataset with over 2000 tenants, peak-to-average memory demand can vary by two orders of magnitude within minutes, resulting in performance degradation and resource inefficiency in job-level allocations.

Isolation. The second requirement is the isolation between different compute tasks. Since multiple computing threads can be using the same disaggregated memory pool, it's essential to multiplex between applications to improve resource efficiency but at the same time keep the memory of different threads isolated from each other, which means that the memory usage of a particular application should not affect other existing applications. The number of tasks reading and writing to the shared disaggregated memory can change rapidly in serverless analytics which makes the problem even more severe.

Lifetime management. Decoupling compute tasks from their intermediate storage means that the tasks can fail independent of the intermediate data, therefore we need mechanisms for explicity lifetime management of intermediate data.

Data repartitioning. Decoupling tasks from their intermediate data also means that data partitioning upon elastic scaling of memory capacity becomes challenging, especially for certain data types used in serverless analytics (e.g. key-value store). If it's the application's responsibility to perform such repartitioning, it will involve large network transfers betweem compute tasks and the far memory system and massive read/write operations every time the capacity is scaled. What's more, the application need to implement different partitioning strategies for different kind of data structures used. Therefore, new mechansims to efficiently enable data partitioning within the far memory system is essential.

We present Jiffy, an elastic disaggregated-memory system for stateful serverless analytics. Jiffy allocates memory resources at the granularity of small fixed-size memory blocks - multiple memory blocks store intermediate data for individual tasks within a job. Jiffy design is motivated by virtual memory design in operating systems that also does memory allocation to individual process at the granularity of fixed-size memory blocks(pages). Jiffy adapts this design to stateful serverless

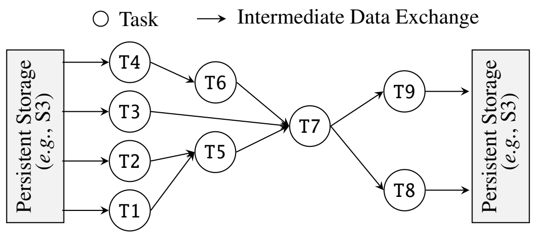

○ Task
→ Intermediate Data Exchange
Persistent Storage (e.g., S3)
T4
T3
T2
T1
T6
T5
T7
T9
T8
Persistent Storage (e.g., S3)

Fig. 2.2: **Execution DAG example for a typical analytics job.** Intermediate data exchange across tasks occurs via Jiffy.

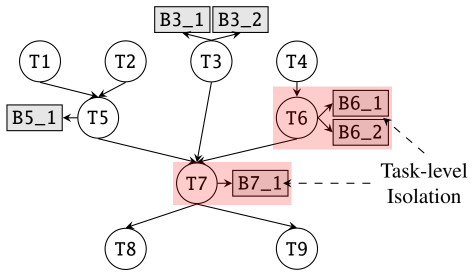


Fig. 2.3: **Hierarchical addressing** for the job in Figure 2.2. Jiffy provides task-level resource isolation for ephemeral storage under each task address-prefix (§2.1.1). Note that block addresses are only assigned to address-prefixes with currently allocated blocks; for tasks `T1`, `T2` and `T4`, blocks are directly read from persistent storage and not stored in Jiffy.

analytics. Performing resource allocation at the granularity of small memory blocks allows Jiffy to elastically scale memory resources allocated to individual jobs without a priori knowledge of intermediate data sizes and to meet the instantaneous job demands at seconds timescales. As a result, Jiffy can efficiently multiplex the available faster memory capacity across concurrently running jobs, thus minimizing the overheads of reads and writes to significantly slower secondary storage (e.g., S3 or disaggregated storage)

2.1 Jiffy Design

This section explains how Jiffy uses hierarchical addressing, intermediate data lifetime management, and flexible data repartitioning to meet these requirements. We illustrate this with Figure 2.2, which depicts the execution plan of a typical analytics job. The plan is represented as a directed acyclic graph (DAG), where nodes are computation tasks (implemented as serverless functions[2]), and edges represent intermediate data exchanged via Jiffy.

2. Functions refer to basic computation units in serverless architectures, such as Amazon Lambdas [58], Google Functions [59], and Azure Functions [60]

2.1.1 Hierarchical Addressing

Analytics jobs are often structured as multiple stages or a directed acyclic graph (DAG). In serverless analytics, where compute elasticity is key, each job can run tens to thousands of tasks [39–53]. Fine-grained resource allocation requires efficient mapping between tasks and their storage blocks, especially with rapidly changing task concurrency. High concurrency demands task-level isolation, ensuring that task arrival or departure doesn't affect others, avoiding performance degradation.

Jiffy adopts a hierarchical addressing mechanism, inspired by the Internet's IP addressing, to maintain task-to-storage mappings and ensure task-level isolation. Jiffy organizes intermediate data in a virtual address hierarchy based on task dependencies in the DAG. Internal nodes represent tasks, and leaf nodes represent Jiffy blocks storing data. Block addresses are defined by the hierarchy path, with task-generated prefixes. Dependencies between tasks are captured by edges between nodes. Jiffy builds this hierarchy from execution plans (e.g., AWS Step Functions, Azure Durable Functions) or dynamically deduces it via the Jiffy API, supporting dynamic query plans without predefined DAGs.

Example. Figure 2.3 illustrates the address hierarchy for the job in Figure 2.2. Internal nodes `T1`-`T9` represent tasks in the DAG, while leaf nodes `B3_1`, `B3_2`, etc., represent data blocks allocated by Jiffy for intermediate data storage. Edges like (`T1`, `T5`) and (`T2`, `T5`) indicate that `T5` depends on the intermediate data from both `T1` and `T2`. The full address of block `B6_2` under `T6` is `T4.T6.B6_2`, with `T4.T6` identifying all blocks under `T6`. Jiffy constructs the address hierarchy either using the execution plan from Figure 2.2 or deduces it dynamically. For instance, Jiffy can infer that since `T7`'s sub-tasks access data from `T3`, `T5`, and `T6`, these tasks must be its parents in the hierarchy.

By organizing intermediate data in a hierarchy, Jiffy manages resource allocation per address prefix. If one prefix spills to persistent storage (via Pocket), it doesn't affect others. Blocks remain assigned until reclaimed or leases expire (§2.1.2), ensuring task-level isolation regardless of churn. Like virtual memory isolating processes, Jiffy uses hierarchical addressing to isolate tasks based on the job's structure.

Two design considerations arise: (1) Jiffy's fine-grained allocation is independent of fairness policies, which can be layered on top, and (2) address translation from virtual to physical storage happens at a centralized metadata server (like Pocket [42]), scaling to arbitrary DAG sizes. Despite

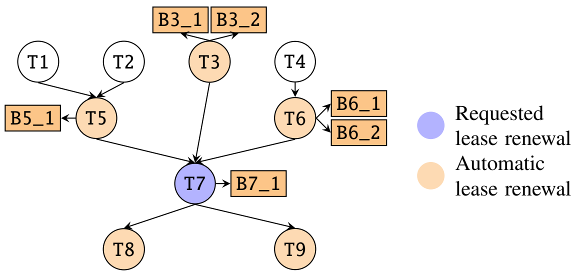


Fig. 2.4: **Lease Renewal via Address Hierarchy.** Hierarchical addressing simplifies lease renewal in Jiffy (§2.1.2), since lease renewal for an address-prefix automatically implies renewals for all parent and descendent address-prefixes in the hierarchy.

added complexity, Jiffy scales to ∼45K requests/sec/core, sufficient for most deployments.

Block sizing. Jiffy balances metadata storage and memory use with block sizes, like 128MB in HDFS [61]. Larger blocks reduce metadata but risk fragmentation, while smaller blocks improve utilization at a metadata cost. Jiffy mitigates this via fine-grained access and data repartitioning.

Isolation granularity. Task-level isolation, where nodes in the hierarchy map to tasks, is default, but finer or coarser isolation (e.g., table-level or stage-level) can be configured via the Jiffy API.

2.1.2 Data Lifetime Management

Existing ephemeral storage systems manage data at the job level, reclaiming storage when the job deregisters. In serverless analytics, decoupled task execution and storage can lead to orphaned data. Jiffy addresses this by integrating lease management [62–64] with hierarchical addressing for task-level data management. Each address prefix has a lease, and data is retained as long as the lease is renewed. Serverless platforms can trigger lease renewals during task monitoring.

Using the DAG hierarchy, Jiffy renews leases for dependent and ancestor tasks automatically, reducing overhead while preventing orphaned data. This strikes a balance between age-based eviction and explicit resource management, ensuring efficient reassignment of resources upon task or job failure.

Example. In Figure 2.2, task `T7`'s job periodically renews the lease for the prefix `T4.T6.T7`[3]. Renewing `T7`'s lease also renews those for parent tasks (`T3`, `T5`, `T6`) and descendants (`T8`, `T9`), as shown in Figure 2.4. This ensures that both parent and descendant tasks' data remain accessible.

3. Task `T7` has four address prefixes; the job can renew any.

Lease duration. Lease duration trades off control plane bandwidth with system utilization. Longer leases reduce network traffic but may delay resource reclamation. Configuring lease durations, as studied in prior work [62, 63], allows Jiffy to meet specific deployment goals.

2.1.3 Flexible Data Repartitioning

Decoupling compute tasks from their intermediate data in serverless analytics introduces challenges in achieving fine-grained elasticity for ephemeral storage. Specifically, when storage is allocated or deallocated for a task, the intermediate data must be efficiently repartitioned across available blocks. However, due to the decoupling of compute from storage and the large number of concurrent tasks, this repartitioning should not be managed by the application itself. For instance, many serverless analytics systems [41, 42] rely on key-value stores for intermediate data. If compute tasks were responsible for repartitioning during memory scaling, they would need to read key-value pairs over the network, compute new partitions based on updated memory, and write the data back to the store—resulting in significant network latency and bandwidth overhead.

Jiffy supports various data structures commonly used in data analytics frameworks, including files [39,44,50–52], key-value pairs [40–42,45,47,49,53], and queues [43,46]. Analytics jobs using these structures can delegate intermediate data repartitioning to Jiffy during resource allocation or deallocation. Each block in a Jiffy data structure tracks its own memory usage. When usage exceeds a predefined threshold, Jiffy allocates a new block to the corresponding address-prefix[4]. The overloaded block triggers a data-specific repartitioning process, moving some of its data to the newly allocated block. Conversely, when block usage drops below a low threshold, Jiffy merges it with another low-usage block before deallocating the unused block. By allowing the block itself, rather than the compute task, to handle repartitioning, Jiffy minimizes network and compute overhead for the task. Repartitioning is done asynchronously, allowing data access to continue with minimal impact on performance.

Jiffy's supported data structures enable the serverless execution of powerful distributed frameworks like MapReduce [66, 67], Dryad [68], StreamScope [69], and Piccolo [70]. Since files, queues, and key-value stores in analytics frameworks require relatively simple repartitioning (unlike

4. Similar to existing systems [24, 25, 42, 65], Jiffy can scale cluster capacity by adding or removing servers based on free blocks. Here, we focus on fine-grained elasticity.

API Group	Function Signature
	connect(honeycombAddress)

Address Hierarchy	createAddrPrefix(addr, parent, optionalArgs)
	createHierarchy(dag, optionalArgs)
	flushAddrPrefix(addr, externalPath)
	loadAddrPrefix(addr, externalPath)

Lease Operations	leaseDuration = getLeaseDuration(addr)
	renewLease(addr)

Data Structure	ds = initDataStructure(addr, type)
	listener = ds.subscribe(op)
	notif = listener.get(timeout)

Table 2.1: **Jiffy User-facing API**: Functions for connecting, managing address hierarchies, handling leases, and interacting with data structures.

complex structures such as B-trees), serverless applications can leverage Jiffy 's flexible repartitioning mechanism without requiring any modifications.

Thresholds for Elastic Scaling. In Jiffy, the high and low thresholds play a crucial role in balancing network bandwidth usage, task performance, and overall system utilization. Setting thresholds too high or too low can impact elastic scaling behavior—if scaling is triggered too infrequently, it may reduce network traffic, but it can also result in inefficient block utilization, such as numerous underutilized blocks. The optimal values for these thresholds depend heavily on the specific workload characteristics, as highlighted in previous studies [71,72]. To accommodate diverse workloads, Jiffy makes these thresholds fully configurable, allowing users to adjust them to suit their performance and efficiency needs.

2.2 Jiffy Implementation

Jiffy builds on Pocket, inheriting its scalable and fault-tolerant metadata plane, multi-tiered data storage, system-wide capacity scaling, and analytics execution model. However, Jiffy introduces hierarchical addressing, lease management, and efficient data repartitioning to address the unique challenges of serverless environments. Below, we describe the Jiffy interface and implementation, highlighting these key features.

2.2.1 Jiffy Interface

We describe the Jiffy interface in terms of its user-facing API (Table 2.1) and internal API (Figure 2.5).

User-facing API. The user-facing interface (Table 2.1) is centered around two core abstractions: *hierarchical addresses* and *data structures*. Jobs can create a new address-prefix using `createAddrPrefix`, specifying the parent prefix along with optional parameters such as initial capacity. The `createHierarchy` function generates a complete hierarchy from an execution plan (DAG), while `flush` and `load` facilitate persisting and retrieving address-prefix data from external storage (e.g., S3). Three built-in data structures can be initialized for an address-prefix via `initDataStructure`, and new structures can be defined using the internal API.

Similar to existing systems [24, 73], Jiffy 's data structures provide a notification interface, allowing tasks that consume intermediate data to be informed when new data is available. For example, a task can `subscribe` to write operations on its parent task's data structure and receive a `listener` handle. When data is written, Jiffy asynchronously notifies the `listener`, which the task can access via `listener.get()`.

```
block = ds.getBlock(op, args) // Get block
block.writeOp(args) // Perform write
data = block.readOp(args) // Perform read
block.deleteOp(args) // Perform delete
```

Fig. 2.5: **Jiffy Internal API.** The block interface is used internally in Jiffy to implement the data structure APIs.

Internal API. The data layout within Jiffy blocks is tailored to the specific data structure that owns it. Therefore, Jiffy blocks expose a set of data structure-specific *operators* (Figure 2.5) that define how requests are *routed* across blocks and how data is *accessed* or *modified*. These operators are used internally by Jiffy for its built-in data structures (§2.2.3) and are not directly exposed to jobs.

The `getBlock` operator determines the target block for an operation based on the operation type and its arguments (e.g., key hashes for a KV-store), returning a handle to the appropriate block. Each Jiffy block provides `writeOp`, `readOp`, and `deleteOp` operators, which implement data structure-specific access logic (e.g., `get`, `put`, and `delete` in a KV-store). Jiffy executes these operators *atomically* using sequence numbers but does not support atomic transactions across multiple operators.

Table 2.2: **Jiffy Data Structure Implementations**. See §**??** for details.

	Data Structure	writeOp	readOp	deleteOp	getBlock	repartition
Built-in	File (§2.2.3)	`write`	`read`	-	File offsets.	✗
	FIFO Queue (§2.2.3)	`enqueue`	`dequeue`		Tail/head.	✗
	KV-Store (§2.2.3)	`put`	`get`	`delete`	Key hash.	✓
Custom data structures						

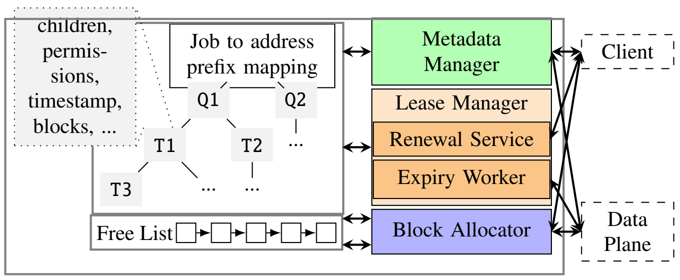


Fig. 2.6: **Jiffy controller.** See §2.2.2 for details.

2.2.2 System Implementation

Since Jiffy design builds on Pocket, its high-level design components are also similar, except for one difference: Jiffy combines the control and metadata planes into a unified control plane. We found this design choice allowed us to significantly simplify interactions between the control and metadata components, without affecting their performance. While this does couple their fault-domains, standard fault-tolerance mechanisms (*e.g.*, the one outlined in [42]) are still applicable to the unified control plane.

Control plane. The Jiffy controller (Figure 2.6) maintains two pieces of system-wide state. First, it stores a *free block list*, which lists the set of blocks that have not been allocated to any job yet, along with their corresponding physical server addresses. Second, it stores an address hierarchy per-job, where each node in the hierarchy stores variety of metadata for its address-prefix, including access permissions (for enforcing access control), timestamps (for lease renewal), a block-map (to locate the blocks associated with the address-prefix in the data plane), along with metadata to identify the data structure associated with the address-prefix and how data is partitioned across its blocks. The mapping between jobIDs (which uniquely identify jobs) and their address hierarchies is stored in a hash-table at the controller.

While the block allocator and metadata manager are similar to their counterparts in Pocket, the lease manager implements lifetime management in Jiffy. It comprises a lease renewal service that

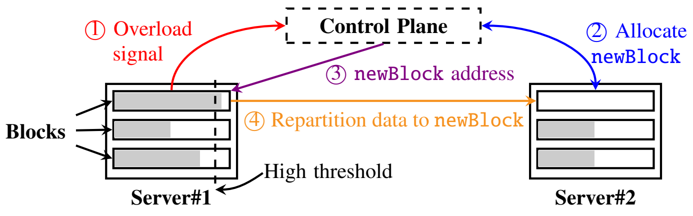


Fig. 2.7: **Data repartitioning on scaling up capacity.** Scaling down capacity employs a similar approach (§2.2.2).

listens for renewal requests from jobs and updates the lease renewal timestamp of relevant nodes in its address hierarchy, and a lease expiry worker that periodically traverses all address hierarchies, marking nodes with timestamps older than the associated lease period as expired. Finally, Jiffy adopts mechanisms from Pocket to facilitate control plane scaling and fault tolerance; we refer the reader to [42] for details.

Data plane. Jiffy data plane is responsible for two main tasks: providing jobs with efficient, data-structure specific atomic access to data, and repartitioning data across blocks allocated by the control plane during resource scaling. It partitions the resources in a pool of storage servers across fixed sized blocks. Each storage server maintains, for the blocks managed by it, a mapping from unique blockIDs to pointers to raw storage allocated to the blocks, along with two additional metadata: data structure-specific operator implementations as described in §**??**, and a subscription map that maps data structure operations to client handles that have subscribed to receive notifications for that operation.

Data repartitioning for a Jiffy data structure is implemented as follows: when a block's usage grows above the high threshold, the block sends a signal to the control plane, which, in turn, allocates a new block to the address-prefix and responds to the overloaded block with its location. The overloaded block then repartitions and moves part of its data to the new block (see Figure 2.7); a similar mechanism is used when the block's usage falls below the low threshold.

For applications that require fault tolerance and persistence for their intermediate data, Jiffy supports chain replication [74] at block granularity, and synchronously persisting data to external stores (*e.g.*, S3) at address-prefix granularity.

2.2.3 Programming Models on Jiffy

We now describe how Jiffy's built-in data structures (Table 2.2) enable various distributed programming frameworks on serverless platforms (§2.2.3-§2.2.3).

Map-Reduce Model

A Map-Reduce (MR) program [66] consists of map functions that process input key-value (KV) pairs to generate intermediate KV pairs, and reduce functions that merge all intermediate values for the same intermediate key. MR frameworks [66, 67, 75] parallelize map and reduce functions across multiple workers. Data exchange between map and reduce workers occurs via a shuffle phase, where intermediate KV pairs are distributed to ensure that values with the same key are routed to the same worker.

In Jiffy, MR executes map/reduce tasks as serverless tasks. A master process launches, tracks, and manages task failures across MR jobs. Jiffy stores intermediate KV pairs in multiple shuffle files, each containing a partitioned subset of KV pairs from all map tasks. Since multiple map tasks may write to the same shuffle file, Jiffy's strong consistency semantics ensure correctness. The master process also handles explicit lease renewals. We now describe Jiffy files in more detail.

Jiffy Files. A Jiffy file consists of multiple blocks, each storing a fixed-sized chunk of the file. The controller manages the mapping between blocks and file offset ranges at the metadata manager, and clients cache this mapping when accessing the file. The mapping is updated whenever the number of blocks allocated to the file scales. The `getBlock` operator forwards requests to the correct file block based on the request's offset range. Files support sequential `reads` and append-only `writes`. For random access, files support `seek` with arbitrary offsets, using the offset to locate the corresponding block. Since files are append-only, blocks are only added and do not require repartitioning when new blocks are added.

Dataflow and Streaming Dataflow Models

In the dataflow model, applications describe their communication patterns using directed acyclic graphs (DAGs), where DAG vertices represent computations, and data channels form directed edges between them. We refer to Dryad [68] as a reference dataflow execution engine, where channels can be files, shared memory FIFO queues, etc. The Dryad runtime schedules DAG vertices based on their dataflow dependencies: a vertex is scheduled once all its input channels are ready. A file channel is ready if its data has been fully written, while a queue is ready if it contains any data. Streaming dataflow [69] adopts a similar approach but operates on continuous event streams.

On Jiffy, each DAG vertex corresponds to a serverless task, with a master process managing vertex scheduling, fault tolerance, and lease renewals. Jiffy uses FIFO queues and files as data channels. Queue-based channels are considered ready as long as a vertex is writing to them, and Jiffy allows downstream tasks to efficiently detect item availability via notifications, described below.

Jiffy Queues. The FIFO queue in Jiffy is implemented as a growing linked-list of blocks, each storing multiple data items and a pointer to the next block. Queue size can be limited by setting a `maxQueueLength`. The controller manages only the head and tail blocks of the queue, and clients cache and update this information when blocks are added or removed. The FIFO queue supports `enqueue` and `dequeue` operations for adding and removing items. The `getBlock` operator routes these operations to the current tail and head blocks, respectively. Unlike other structures, queues do not require repartitioning. The FIFO queue uses Jiffy 's notification system to asynchronously detect when there is space to add items or when data is available for consumption through subscriptions to `enqueue` and `dequeue` events.

Piccolo

Piccolo [70] is a data-centric programming model that allows distributed machines to share mutable state. Piccolo kernel functions define sequential application logic, while sharing state with concurrent kernel functions via a KV interface. Centralized control functions create and coordinate both shared KV stores and kernel instances. Concurrent updates to the same key are resolved using user-defined accumulators.

On Jiffy, Piccolo kernel functions execute across serverless tasks, while control tasks run on a centralized master. Shared state is stored in Jiffy 's KV-store data structures (described below), which may be created per kernel function or shared across multiple functions as needed. The master periodically renews leases for Jiffy KV-stores, and, like Piccolo, Jiffy checkpoints KV-stores by flushing them to external storage.

Jiffy KV-store. The Jiffy KV-store hashes each key into one of H hash slots, where $H = 1024$ by default. The KV-store shards KV pairs across multiple Jiffy blocks, with each block responsible for one or more hash slots. A hash slot is fully contained within a single block. The controller manages the mapping between blocks and their corresponding hash slots, and this mapping is cached at the client and updated during scaling. Each block stores KV pairs as a hash table. The KV-store supports standard `get`, `put`, and `delete` operations via `readOp`, `writeOp`, and `deleteOp` operators. The `getBlock` operator routes requests to blocks based on key hashes.

Unlike files and queues, the KV-store requires repartitioning when blocks are added or removed. When a block becomes nearly full, Jiffy splits half of its hash slots into a new block, moves the relevant KV pairs, and updates the mapping at the controller. Similarly, when a block is underutilized, its hash slots are merged with another block.

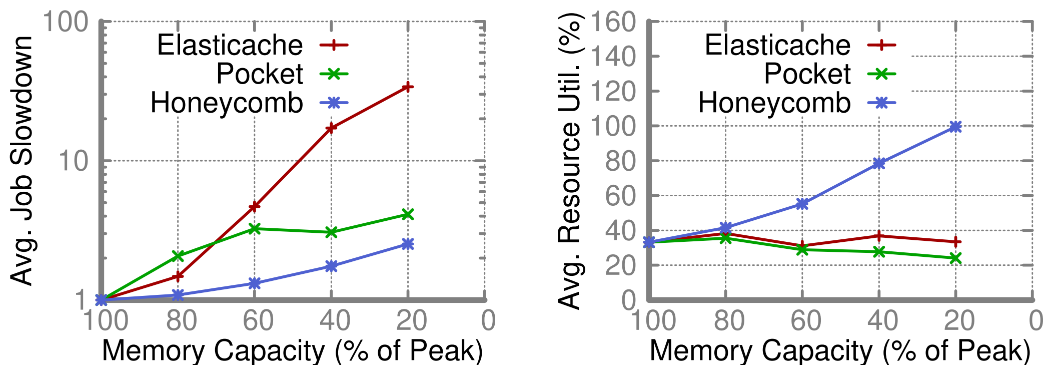
2.3 Evaluation

Jiffy is implemented in 25K lines of C++, with client libraries in C++, Python, and Java (each around 1K LOC), in addition to the original Pocket codebase. In this section, we evaluate Jiffy to showcase its benefits (§2.3.1, §2.3.2) and analyze the contributions of individual Jiffy mechanisms to overall performance (§2.3.1). Lastly, we assess Jiffy 's controller overheads in §A.1.1.

Experimental setup. Unless specified otherwise, each intermediate storage system in our experiments is deployed across 10 m4.16xlarge EC2 [76] instances, while serverless applications are hosted on AWS Lambda [76]. Since Jiffy builds on Pocket's design, it supports adding new instances to increase overall system capacity. However, our experiments do not evaluate the overheads of scaling system capacity, as this is orthogonal to Jiffy 's focus. Instead, we concentrate on multiplexing the available storage capacity for higher utilization, reducing the need to add more resources. Jiffy employs 128MB blocks, a 1-second lease duration, and thresholds of 5% (low) and 95% (high) for data repartitioning.

2.3.1 Benefits of Jiffy

Jiffy enables fine-grained resource allocation for serverless analytics. We demonstrate its impact on job performance and resource utilization across approximately 50,000 jobs from 100 randomly



(a) Job performance (b) Resource utilization

Fig. 2.8: **Fine-grained task-level elasticity in Jiffy** enables (a) better job performance, and (b) higher resource utilization under constrained capacity. In (a), the slowdown is computed relative to the job completion time with 100% capacity (for this data point, Elasticache performance was 30% worse than Pocket, and Pocket performance was 5% worse than Jiffy). See §2.3.1 for details.

selected tenants over a 5-hour window in the Snowflake workload[5] [39].

We compare Jiffy (using the MR programming model, §2.2.3) with Elasticache [65] and Pocket [42]. Elasticache provisions resources for *all* jobs, and if capacity is insufficient, jobs must spill data to external stores like S3 [77]. Pocket, however, reserves and reclaims resources at a *job* granularity, spilling data to SSD if DRAM capacity is insufficient. Pocket's utilization can be lower than Elasticache, as it provisions separately for the peak demand of each job, which sacrifices overall utilization. To ensure a fair comparison, Pocket's control and metadata services are colocated on the same server, similar to Jiffy 's unified control plane.

Impact of fine-grained elasticity on job performance. We examine job performance under constrained intermediate storage capacity in the Snowflake workload. Figure 2.8(a) shows the average job slowdown as capacity is reduced to a fraction of peak utilization. With Elasticache, performance drops sharply when data exceeds capacity, leading to a 34× slowdown at 20% capacity due to reliance on S3. Pocket experiences a $>$ 4.1× slowdown at 20% capacity as it spills data to SSD. In contrast, Jiffy 's task-level elasticity and lease-based storage reclamation reduce data spilling, resulting in a much lower slowdown ($<$ 2.5× at 20% capacity). This is because Jiffy multiplexes capacity more efficiently across jobs, minimizing reliance on slower storage tiers.

Impact of fine-grained elasticity on resource utilization. Figure 2.8(b) shows resource utilization under constrained capacity. While Elasticache and Pocket see reduced or stagnant utilization

5. We did not evaluate the full 14-day window with $>$ 2000 tenants due to intractable cost overheads.

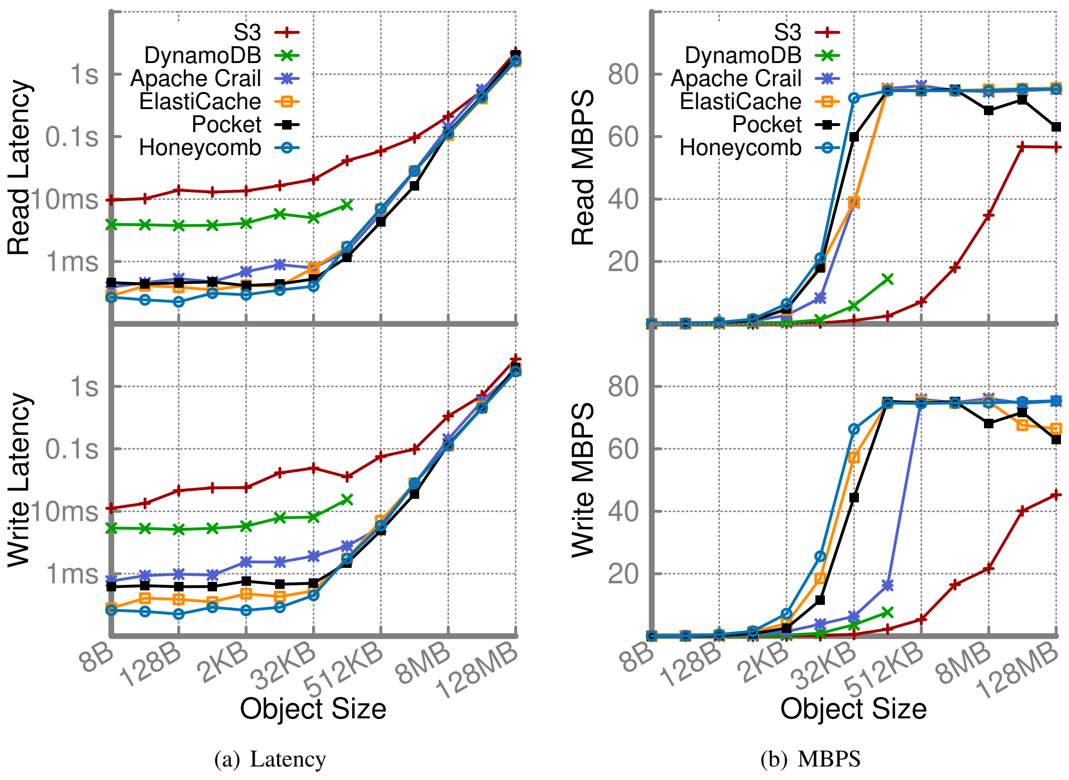


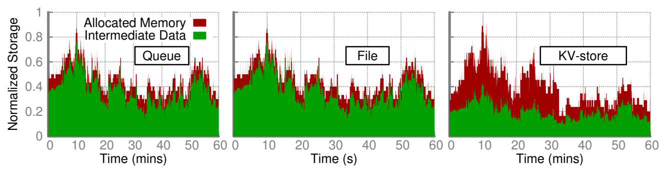
Fig. 2.9: **Jiffy performance comparison with existing storage systems (§2.3.2).** Despite providing the additional benefits demonstrated in §2.3.1, Jiffy performs as well as or outperforms state-of-the-art storage systems for serverless analytics.

as capacity is constrained, Jiffy 's utilization *improves*. Elasticache and Pocket allocate capacity at a job or coarser granularity, wasting unused resources regardless of total system capacity. In contrast, Jiffy 's fine-grained elasticity and lease-based reclamation allow it to multiplex capacity more effectively, reducing SSD spillover and improving performance as shown in Figure 2.8(a).

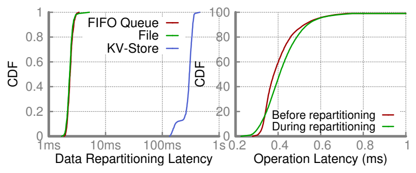
2.3.2 Performance Benchmarks for Six Systems

We now compare Jiffy 's performance (using its KV-Store data structure) against five state-of-the-art systems commonly used for intermediate data storage in serverless analytics: S3, DynamoDB, Elasticache, Apache Crail, and Pocket. Since only a subset of these systems support request pipelining, we disable pipelining across all of them for consistency.

To measure latency and throughput, we profiled synchronous operations issued from an AWS Lambda instance using a single-threaded client. Figure 2.9 shows that in-memory data stores like Elasticache, Pocket, and Apache Crail achieve low latency (sub-millisecond) and high throughput.



(a) Efficient lifetime management



(b) Efficient data repartitioning

Fig. 2.10: **Jiffy data lifetime-management and data repartitioning.** (a) Jiffy provides fine-grained elasticity through lease-based lifetime management for its built-in data structures: FIFO Queue (left), File (center), and KV-store (right). It efficiently reclaims resources from tasks once their leases expire. (b) Jiffy enables efficient data repartitioning as allocations scale up, with repartitioning for a single block completing within 2-500ms (left). Additionally, the latency for 100KB `get` operations is minimally affected during KV-store repartitioning. Note: plots in (a) and (b) share a common y-axis; the x-axis for (c, left) is in log scale.

In contrast, persistent data stores like S3 and DynamoDB exhibit significantly higher latencies and lower throughput; note that DynamoDB only supports objects up to 128KB. Jiffy matches the performance of these in-memory data stores while also providing the additional benefits discussed in §2.3.1.

2.3.3 Understanding Jiffy Benefits

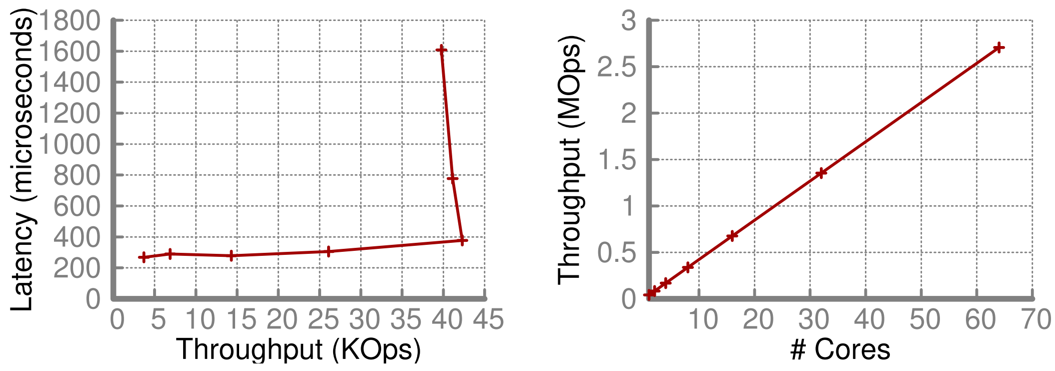
Figure 2.8 demonstrates how Jiffy's fine-grained elasticity provides performance and resource utilization advantages over other state-of-the-art systems. This elasticity is achieved through hierarchical virtual addressing, flexible data lifetime management, and data repartitioning. In this section, we isolate and evaluate the impact of these mechanisms.

Fine-grained elasticity via data lifetime management. Unlike traditional storage systems, Jiffy 's lease-based data lifetime management enables the reclamation of unused resources, reallocating them to jobs in need. Coupled with fine-grained resource allocations and efficient data repartitioning, this enables elasticity for serverless jobs. To evaluate this, we examine storage allocation across various Jiffy data structures (Figure 2.10(a)) using the Snowflake workload from Figure 2.1.

FIFO queue and file data structures exhibit seamless elasticity as intermediate data is written to them, as they do not require repartitioning. The allocated capacity slightly exceeds the intermediate data size, accounting for block metadata (e.g., object metadata for FIFO queue items) and unused space in head/tail blocks. For the KV-store, inserted keys are sampled from a Zipf distribution since the Snowflake dataset lacks access patterns. Due to the skew, some Jiffy blocks receive most key-value pairs and frequently split when their capacity grows too high, leading to higher allocated capacity. However, Jiffy 's lease mechanism quickly reclaims resources after their utility ends, ensuring that overheads are temporary.

Efficient elastic scaling via flexible data repartitioning. A key factor in Jiffy 's elasticity is its efficient data repartitioning. Figure 2.10(b) shows the CDF of repartitioning latency per block across the three data structures under the Snowflake workload. The latency includes the time from detecting an overloaded/underloaded block to the completion of repartitioning. Storage servers take ∼1-1.5ms to connect to the controller, with two round trips (100-200μs in EC2) to trigger block allocation/reclamation and update partitioning metadata. Unlike FIFO Queue and File, KV-Store requires data repartitioning across blocks, but since only half the block capacity (∼64MB) is moved, Jiffy completes repartitioning in a few hundred milliseconds over 10Gbps links, achieving block-level repartitioning with low latency (2-500ms).

Importantly, Jiffy does not block data structure operations during repartitioning. As shown in Figure 2.10(b), the CDF of 100KB `get` operations in the KV-Store before and during scaling remains almost identical, indicating minimal impact on operation latency during scaling.



(a) Controller throughput vs. latency on a single CPU core.

(b) Controller throughput scaling with multiple cores.

Fig. 2.11: **Jiffy controller performance.** Details in Appendix A.1.1.

2.3.4 Controller Overheads

Jiffy introduces several additional components at the controller compared to Pocket, including metadata management, lease management, and handling data repartitioning requests. As a result, its performance is expected to be lower than Pocket's metadata server. However, this is acceptable as long as Jiffy can manage the typical control plane request rates observed in real-world workloads, such as the peak of a few hundred requests per second—including lease renewals—seen in our evaluations and those in [42].

Figure A.1(a) shows the throughput-versus-latency curve for Jiffy controller operations on a single CPU core of an m4.16xlarge EC2 instance. The controller throughput saturates at around 42 KOps with a latency of 370μs. While this is lower than Pocket's throughput (90 KOps per core), it is more than sufficient to handle the control plane load of real-world workloads. Additionally, throughput scales almost linearly with the number of cores, as each core processes requests independently for distinct subsets of virtual address hierarchies (Figure A.1(b)). Finally, the control plane can scale across multiple servers by partitioning the address hierarchies.

Storage overheads. The task-level metadata storage in Honeycomb has a minimal overhead of just 64 bytes of fixed metadata per task and 8 bytes per block. For Jiffy's default 128MB blocks, this results in an insignificant storage overhead ($< 0.00005 - 0.0001\%$ of total storage).

2.4 Related Work

2.5 Summary

In this chapter, we have presented Jiffy, a memory management service designed for disaggregated memory, allocating memory in small fixed-size blocks. These blocks store intermediate data for individual tasks within a job. Inspired by virtual memory systems in operating systems, Jiffy scales memory resources elastically without prior knowledge of data sizes, adapting to job demands in real-time. This approach allows Jiffy to efficiently share fast memory across jobs, reducing reliance on slower secondary storage such as S3 or disaggregated storage.

Chapter 3

Operating System Layer: In-Network Memory Management

In the previous chapter, we explored the design of memory management for disaggregated architectures at the service layer. Specifically, we examined how serverless applications, which are inherently aware of disaggregated memory and compute resources, require explicit memory management for handling intermediate data. Integrating such applications with Jiffy is straightforward, as they can directly benefit from Jiffy's elasticity and lifetime management features. However, general-purpose applications (beyond serverless data analytics) are typically developed without any knowledge of the underlying hardware specifics, such as disaggregated resources. As a result, integrating these applications with external memory services (e.g., Jiffy) often necessitates significant code modifications to accommodate their APIs, which may present an undesirable burden for developers.

In traditional monolithic architectures, memory management is typically handled by the operating system (OS), which manages virtual and physical pages, performs memory address translation, enforces memory protection, and provides a simple interface to user applications. This abstraction hides the hardware details, thereby easing the burden of memory management for developers. If the OS is made aware of disaggregated architectures and can transparently manage memory, it would be possible to migrate existing applications to these architectures without requiring any code modifications.

The fundamental distinction between performing memory management at the OS layer versus the service layer lies in the scope and specificity of the functionality provided. The OS must offer highly general functionality that applies to all applications, while the service layer can afford to provide more specialized features tailored to specific application types (e.g., lifetime management in Jiffy). A key question that arises is where the OS should be situated within the disaggregated architecture. Unlike monolithic architectures, where the OS resides directly on each server, disaggregated architectures lack a single, centralized server. We observe that the network interconnect (e.g., Ethernet) presents a promising point for implementing OS-level memory management (3.1). However, even if memory management functionalities are implemented transparently within the OS, their performance may vary depending on the application type due to fundamental differences in how resources are organized (3.2).

In this chapter, we shift our focus to embedding memory management functionality directly within the OS. We first introduce MIND, an in-network OS design that enables transparent memory management for disaggregated resources. Following this, we present PULSE, an in-network near-memory accelerator designed to optimize performance for pointer traversal workloads.

3.1 MIND: In-Network Memory Management for Disaggregated Data Centers

Implementing memory management at the OS layer in disaggregated architectures poses three significant challenges. First, remote memory access requires low latency and high throughput, with targets of 10 μs latency and 100 Gbps bandwidth per compute blade [2,3,31,32]. Second, both compute and memory resources must scale elastically to meet varying workloads. Finally, widespread adoption of disaggregated memory necessitates support for unmodified applications, minimizing the need for developers to rewrite code.

We introduce MIND, the first memory management system designed for rack-scale disaggregated memory, addressing these challenges by embedding the memory management module (logic and metadata for memory management) directly within the network fabric and leveraging programmable network switches [78, 79].

The placement of MIND's memory management *within the network fabric* is motivated by three

Table 3.1: **Parallels between memory & networking primitives.**

Virtual Memory	$\Longleftrightarrow$	**Networking**
Memory allocation		IP assignment
Address translation		IP forwarding
Memory protection		Access control
Cache invalidations		Multicast

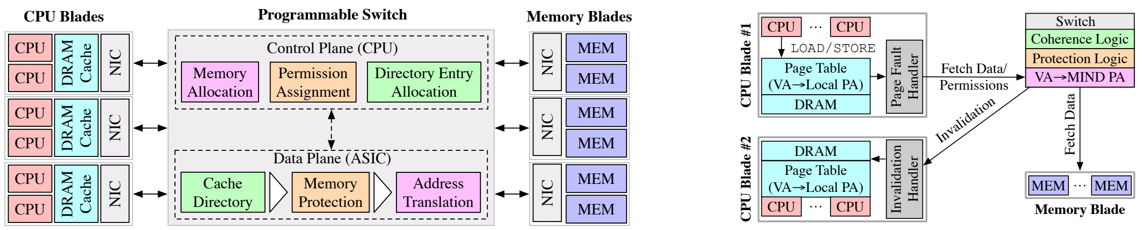


Fig. 3.1: **(left) High-level MIND architecture, and, (right) data flow for memory accesses in MIND.**

key factors: (1) its central location provides a global view of the system, enabling direct memory access without requiring metadata coherence, (2) virtual memory access bears a structural similarity to network address access 3.1, and programmable network switches [78] are capable of executing at line rate, making them suitable for implementing memory management logic, and (3) incorporating cache coherence logic into the network fabric helps reduce latency and bandwidth overhead.

MIND provides a *transparent virtual memory* abstraction to applications, functioning similarly to traditional OS mechanisms. It intercepts memory allocations on CPU blades and performs memory operations via RDMA, using a switch-based MMU for managing cache coherence. Memory blades store pages and directly handle RDMA requests, enabling true hardware disaggregation.

Figure 3.1(left) presents an overview of MIND's design, while Figure 3.1(right) illustrates its memory access flow. CPU blades run user processes and utilize local DRAM as a cache. Memory allocations and deallocations are intercepted and forwarded to the switch control plane, which manages memory allocation and access permissions. All memory operations are handled by the CPU cache, with virtual addresses translated locally. If a requested page is not cached, a page fault triggers an RDMA request to fetch it from the memory blades. Coherence updates may also trigger page faults, which are handled by the switch.

Since CPU blades do not maintain memory metadata, RDMA requests operate solely on virtual addresses. The switch's data plane intercepts these requests, managing cache coherence, permission verification, and address translation. If no cache contains the requested page, the switch forwards the request to the appropriate memory blade. MIND relies on one-sided RDMA, removing the need for CPUs on memory blades and paving the way for complete hardware disaggregation.

3.2 Need for PULSE

Recent disaggregated systems [1, 2] use small DRAM caches on CPU nodes while accessing memory across network-attached memory nodes with large DRAM pools (Fig. 3.2 (top)). However, limited bandwidth and latency of network-attached memory remain a challenge, constrained by the speed of light. Even with near-terabit links and RDMA [80], remote memory is much slower than local memory [3]. CXL interconnects [10] show similar patterns, with 300 ns latency compared to 10–20 ns for L3 cache [37].

CPU caches can reduce average memory access latency, but their effectiveness is limited by data locality and cache size. Remote memory access is unavoidable for pointer-heavy applications, such as database index lookups [81–91] and graph analytics [92–95] (Fig. 3.3). Memory-intensive applications [67, 96–101] often require traversing linked structures like lists, hash tables, trees, and graphs. Despite large memory pools in disaggregated architectures, network pointer traversals remain slow [3]. Recent systems [1–3, 31, 32] mitigate this by caching hot data in CPU DRAM, but pointer traversals still suffer, as we demonstrate next.

Pointer traversals in real-world workloads. Studies [67, 94, 102–106] show that cloud applications spend 21% to 97% of their execution time on pointer traversals. We analyzed three representative cloud applications — a WebService frontend [28], WiredTiger indexing [107], and BTrDB time-series analysis [108] — using swap-based disaggregated memory [32]. Varying the CPU cache size from 6.25% to 100% of the working set, Fig. 3.3(a) shows that (i) significant execution time is spent on pointer traversals (13.6%, 63.7%, and 55.8% respectively, even with full cache), and (ii) traversal time increases as CPU cache size decreases.

Distributed traversals. As application workloads and working-set sizes grow, disaggregated systems allocate memory across multiple memory nodes [1, 2, 31, 32]. To optimize load balancing and utilization, they use fine-grained allocations (e.g., 1 GB in [2], 2 MB in [1]), but this fragments linked structures across memory nodes, leading to frequent distributed traversals.

Fig. 3.3(b) illustrates this for WiredTiger and BTrDB on a setup with 1 compute and 4 memory nodes: over 97% and 75% of requests, respectively, cross memory node boundaries. Fig. 3.3(c)

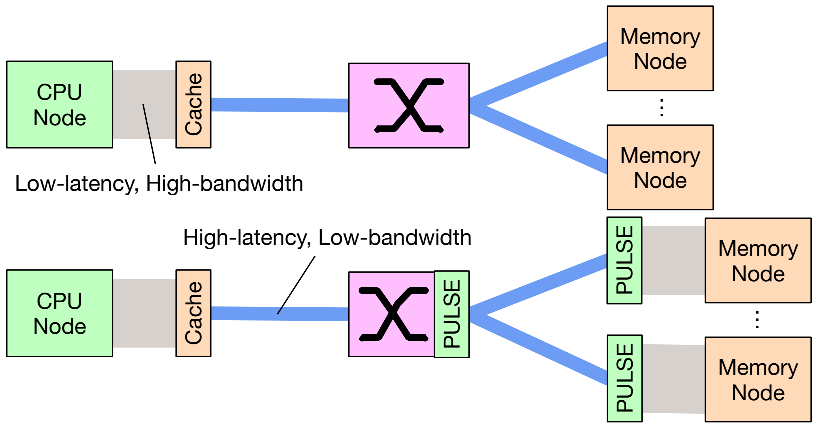
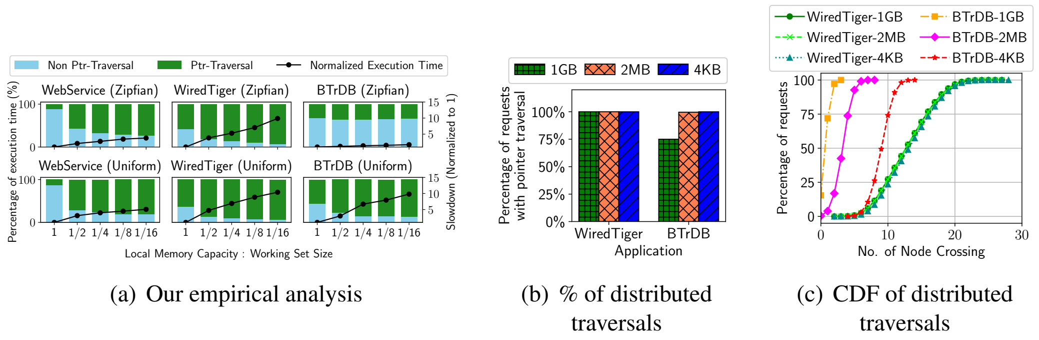


Fig. 3.2: **Need for accelerating pointer traversals.** *(top)* Pointer traversals in disaggregated architectures are limited by slow memory interconnects. *(bottom)* Like CPU caches, we propose a fast, lightweight accelerator for cache-unfriendly pointer traversals in traversal-heavy workloads.

shows the CDF of memory node crossings. While WiredTiger's randomly ordered data requires frequent crossings, BTrDB's time-ordered data confines larger allocations to the same node, reducing crossings. However, smaller allocations, necessary for high utilization, still result in many cross-node traversals.



(a) Our empirical analysis

(b) % of distributed traversals

(c) CDF of distributed traversals

Fig. 3.3: **Time cloud applications spend in pointer traversals.**

Similar to the role of CPU caches in providing quick access to frequently used data, we propose augmenting memory nodes with lightweight, fast processing units that offer high-bandwidth, low-latency access for accelerating pointer traversals (Fig. 3.2 (bottom)). Additionally, the interconnect must enable efficient and scalable traversals across multiple memory nodes to handle large, linked data structures.

We introduce PULSE[1], a distributed framework designed for efficient pointer traversals in rack-

1. **P**rocessing **U**nit for **L**inked **S**tructur**E**s.

scale disaggregated memory systems. PULSE addresses three critical aspects—expressiveness, energy efficiency, and performance—by adopting a novel near-memory processing paradigm. At the heart of PULSE is an expressive iterator interface that provides a unified abstraction for pointer traversals in a variety of applications, including key-value stores [24, 106], databases [86–88, 90, 107], and big-data analytics frameworks [92–95] (§??). This abstraction supports a wide range of traversal-heavy workloads, enabling (i) seamless integration with existing toolchains and (ii) the deployment of hardware accelerators optimized for iterators.

To ensure efficient pointer traversals, PULSE introduces a novel accelerator that decouples the logic and memory pipelines, leveraging the sequential nature of iterator execution (§??). This design enables high memory utilization by balancing memory capacity with fewer logic pipelines. A scheduler distributes the traversal logic across multiple pipelines, employing multiplexing to maximize resource utilization. While our initial implementation of PULSE leverages FPGA-based SmartNICs due to the complexity of ASIC design, the framework is ultimately aimed at an ASIC-based implementation for higher efficiency.

For distributed traversals, PULSE employs a programmable network switch, treating pointer traversals across memory nodes similarly to packet routing (§??). The switch inspects iterator requests and routes them to the appropriate memory node at line rate. We have implemented a prototype using commodity servers, SmartNICs, and a programmable switch, making minimal hardware and software changes to ensure non-invasive deployment in existing infrastructures.

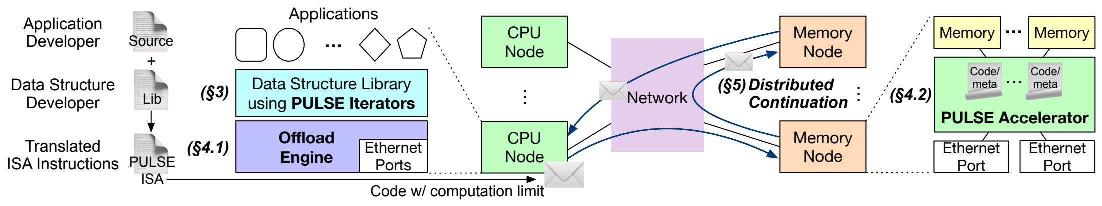


Fig. 3.4: **PULSE Overview.** Developers use PULSE's iterator interface (§3.3.1) to express pointer traversals, translated to PULSE ISA by its dispatch engine (§??). During execution, PULSE accelerator ensures energy efficiency (§3.3.3) and in-network design enable distributed traversals (§3.3.4).

3.3 PULSE Design

PULSE introduces innovations across three key design elements (Fig. 3.4). At its core, PULSE's iterator-based programming model (§**??**) simplifies the process of porting real-world data structure traversals. It supports *stateful* traversals using a *scratchpad*, which allows developers to store and update intermediate states (e.g., aggregators, arrays) during the execution of iterators. This iterator-based approach enables both tractable accelerator design and efficient distributed traversals.

The developer's iterator code is translated into PULSE's ISA (Instruction Set Architecture) for execution by PULSE accelerators (§**??**). The accelerator achieves energy efficiency and high performance by decoupling logic and memory pipelines, with an ISA specifically tailored for the iterator pattern. A specialized scheduler is employed to maximize utilization and performance in this disaggregated architecture.

For scalable distributed pointer traversals, PULSE leverages programmable network switches to reroute requests that cross memory node boundaries (§**??**). Hierarchical address translation is performed *in-network*, with the network switch managing memory node-level translation, while the accelerators at each memory node handle local address translation and memory protection. If a request cannot be handled locally, the accelerator returns it to the switch for rerouting to the appropriate memory node.

Assumptions. PULSE relies on the CPU node for synchronization, requiring the application to explicitly handle locks. Although recent work enables locking mechanisms on NICs [29, 109] and switches [110], these efforts are orthogonal and could be incorporated into PULSE. Additionally, PULSE adopts the caching scheme from prior work [28], maintaining a transparent cache within the data structure library.

3.3.1 PULSE Programming Model

We begin by discussing PULSE's programming model, as a well-designed interface is essential for supporting real-world traversal-heavy applications and for enabling the development of efficient pointer traversal accelerators and distributed mechanisms. PULSE's interface is specifically designed for data structure library developers, allowing them to offload pointer traversals within linked structures. Since the required code modifications are confined to the data structure libraries, applications that use these libraries can operate without any changes.

After analyzing various popular data structures [111–114], we identified a common traversal pattern: (1) initializing a start pointer, (2) iteratively computing the next pointer, and (3) checking a termination condition at the end of each iteration. This pattern closely aligns with the *iterator* design motif, which is widely adopted across different programming languages [113]. Consequently, PULSE adopts the iterator interface as the hardware-software boundary for handling pointer traversals (Listing 3.1).

The interface exposes three user-defined functions: (1) `init()` initializes the starting pointer based on the data structure, (2) `next()` updates the current pointer to the next element, and (3) `end()` checks whether the traversal should terminate. PULSE uses these functions to iteratively execute the traversal through `execute()`. Additionally, we introduce two key features in our iterator abstraction to both enhance and constrain the expressiveness of operations on linked data structures.

```
1  class pulse_iterator {
2      void init(void *) = 0; // Implemented by developer
3      void *next() = 0; // Implemented by developer
4      bool end() = 0; // Implemented by developer
5
6      unsigned char *execute() { // Non-modifiable logic
7        unsigned int num_iter = 0;
8        while (!end() && num_iter++ < MAX_ITER)
9          cur_ptr = next();
10       return scratch_pad;
11     }
12     uintptr_t cur_ptr;
13     unsigned char scratch_pad[MAX_SCRATCHPAD_SIZE];
14 }
```

Listing 3.1: PULSE interface.

Stateful traversals. Pointer traversals in data structures are often stateful, with different types of state depending on the application. For example, in hash table lookups, the state is the search key, whereas in B-Tree summations, a running total is maintained and updated with each value. To support such operations, PULSE iterators include a `scratch_pad` for storing arbitrary state. The state is initialized in `init()`, updated during each iteration in `next()`, and finalized in `end()`. Upon completion, the `execute()` function returns the contents of `scratch_pad` (Line 10), allowing developers to retrieve the result of the traversal.

Bounded computations. PULSE accelerators facilitate lightweight processing for memory-intensive

```
struct node {
  key_type key;
  value_type value;
  struct node *next;
};

value_type find(key_type key) {
  for (struct node *cur_ptr = bucket_ptr(hash(key)); ; cur_ptr = cur_ptr->next) {
    if (key == cur_ptr->key) // Key found
      return cur_ptr->value;
    if (cur_ptr->next == nullptr) // Key not found
      break;
  }
  return KEY_NOT_FOUND;
}
```

Listing 3.2: C++ STL realization for `unordered_map::find()`.

```
class unordered_map_find : pulse_iterator {
  init(void *key) {
    memcpy(scratch_pad, key, sizeof(key_type));
    cur_ptr = bucket_ptr(hash((key_type)*key));
  }

  void* next() { return cur_ptr->next; }

  bool end() {
    key_type key = *((key_type *)scratch_pad);
    if (key == cur_ptr->key) { // Key found
      *((value_type *)scratch_pad) = cur_ptr->value;
      return true;
    }
    if (cur_ptr->next == nullptr) { // Key not found
      *((unsigned int *)scratch_pad) = KEY_NOT_FOUND;
      return true;
    }
    return false;
  }
}
```

Listing 3.3: PULSE realization for `unordered_map::find()`.

operations, ensuring efficient utilization of bandwidth. While `init()` is executed on the CPU, the `next()` and `end()` functions are offloaded to PULSE accelerators. These accelerators enforce two constraints on memory accesses and computations. First, PULSE disallows nondeterministic behavior, such as unbounded loops that cannot be unrolled. Second, `execute()` (as shown in Listing 3.1) limits the maximum number of iterations per request to prevent long-running traversals from monopolizing resources. If this limit is reached, PULSE terminates the traversal and returns the current `scratch_pad` value to the CPU, allowing a new request to continue from the last point.

An illustrative example. To demonstrate, we adapt the `find()` operation from the C++ STL unordered_map to PULSE. Listing 3.2 shows a simplified STL implementation, where the traversal begins by computing a hash function to locate the corresponding hash bucket pointer, followed by iterating through the linked list in that bucket. The traversal ends when the key is found or when the list terminates.

In Listing 3.3, the PULSE iterator implementation is presented. The core logic remains largely unchanged, with modifications made to the init(), next(), and end() functions. The key differences involve how the state (the search key) is passed between these functions and how results (either an error message if the key is not found or its value if it is) are returned via the scratch_pad.

3.3.2 PULSE Dispatch Engine

The dispatch engine is a software framework running on the CPU node, designed for two key purposes. First, it translates the iterator-based pointer traversal code provided by data structure library developers (§**??**) into PULSE's ISA. Second, it determines whether the accelerator can handle the computations required during the traversal, and if so, sends the request to the accelerator at the memory node. If the accelerator is unsuitable, execution proceeds on the CPU with regular remote memory accesses.

Translating iterator code to PULSE ISA. To integrate seamlessly with existing workflows, PULSE plugs into standard compiler toolchains. The dispatch engine generates PULSE ISA instructions using well-established compiler techniques [115]. PULSE's ISA is a streamlined RISC instruction set, designed with only the essential operations for basic processing and memory access, optimizing for simplicity and energy efficiency (Table 3.2). However, there are a few notable features in the adaptation of the iterator code to PULSE's ISA.

First, as discussed in §**??**, PULSE does not support unbounded loops within a single iteration. The ISA only supports conditional jumps that move forward in the code, akin to eBPF programs [116], which allow only forward jumps to prevent infinite execution within the kernel. A backward jump can occur only at the start of the next iteration; PULSE includes a specific NEXT_ITER instruction to explicitly mark this point, enabling the accelerator to begin scheduling the memory pipeline (§3.3.3).

Second, developers can maintain state and return values using the scratch_pad, which has a pre-

Class	Instructions	Description
Memory	`LOAD`, `STORE`	Load/store data from/to address.
ALU	`ADD`, `SUB`, `MUL`, `DIV`, `AND`, `OR`, `NOT`	Standard ALU operations.
Register	`MOVE`	Move data b/w registers.
Branch	`COMPARE` and `JUMP_{EQ, NEQ, LT, ...}`	Compare values & jump ahead based on condition (*e.g.*, equal, less than, etc.).
Terminal	`RETURN`, `NEXT_ITER`	End traversal & return, or start next iteration.

Table 3.2: **PULSE adapts a restricted subset of RISC-V ISA** (§3.3.2).

configured size. PULSE's ISA supports direct register operations on the `scratch_pad` and includes a `RETURN` instruction, which ends the iterator's execution and returns the contents of the `scratch_pad` as the result.

Lastly, we observed that iterator traversal typically involves two main types of operations: fetching data[2] from memory via the `cur_ptr` and processing that data to determine the next pointer or whether the traversal should terminate. If the translation of iterator code to PULSE's ISA is done naively, it may result in multiple redundant loads near the memory location pointed to by `cur_ptr`. For example, in the `unordered_map::find()` implementation in Listing 3.3, references to `cur_ptr->key`, `cur_ptr->value`, and `cur_ptr->next` occur at different points, and each could incur a separate memory load, slowing down execution and wasting memory bandwidth. To address this, PULSE's dispatch engine uses static analysis to infer the range of memory locations accessed relative to `cur_ptr` in the `next()` and `end()` functions and aggregates these accesses into a single large `LOAD` (up to 256 B) at the beginning of each iteration.

Bounding complexity of offloaded code. While PULSE's interface and ISA already limit the *types* of computations that can be performed per iteration, it is also necessary to restrict the *amount* of computation to ensure that the operations offloaded to PULSE accelerators remain memory-centric. To achieve this, PULSE's dispatch engine analyzes the generated ISA for the iterator to estimate the time required for computational logic (t_c) and the time required for the single data load performed at the beginning of each iteration (t_d).

PULSE leverages the known execution time per compute instruction of its accelerators, denoted

2. While this section focuses on data fetches, writing data to memory follows a similar process.

as t_i, to calculate $t_c = t_i \cdot N$, where N represents the number of instructions per iteration. The CPU node will offload the iterator execution only if $t_c \leq \eta \cdot t_d$, where η is a predefined threshold specific to the accelerator. Since PULSE aims to offload only memory-centric operations, $\eta \leq 1$. As discussed in §3.3.3, the choice of η allows PULSE to maximize memory bandwidth utilization and ensures that processing does not become a bottleneck for pointer traversals.

Issuing network requests to the accelerator. Once the dispatch engine decides to offload the iterator execution, it encapsulates the ISA instructions (`code`) along with the initial values of `cur_ptr` and `scratch_pad` (initialized by `init()`) into a network request. This request is then issued to the network, which determines the appropriate memory node to forward the request to (§**??**). To handle potential packet drops, the dispatch engine embeds a unique request identifier (ID) consisting of the CPU node ID and a local request counter within the request packets. The engine also maintains a timer for each request and retransmits requests in case of a timeout.

Practical deployability. PULSE's software stack is easily deployable due to its compatibility with real-world toolchains. Our user library adapts common data structures used in key-value stores [24, 106], databases [86–88, 90, 107], and big-data analytics frameworks [92–95] to PULSE's iterator interface (§**??**). The PULSE dispatch engine is implemented on a low-latency, high-throughput UDP stack based on Intel DPDK [117]. The PULSE compiler adapts the Sparc backend of LLVM [118], which is closely aligned with PULSE's ISA. Additionally, the LLVM frontend applies a set of analysis and optimization passes [119] to enforce PULSE's constraints and semantics: the analysis pass identifies sections of code that require offloading, while the optimization pass translates pointer traversal logic into PULSE ISA.

3.3.3 PULSE Accelerator Design

The accelerator is at the heart of PULSE design and is key to ensuring high performance for iterator executions with high resource and energy efficiency. Our motivation for a new accelerator design stems from two unique properties of iterator executions on linked structures:

- **Property 1:** Each iteration involves two clearly separated but sequentially dependent steps: (i) fetching data from memory via a pointer (*e.g.*, a list or tree node), followed by (ii) executing logic on the fetched data to identify the next pointer. The logic cannot be executed concurrently with or

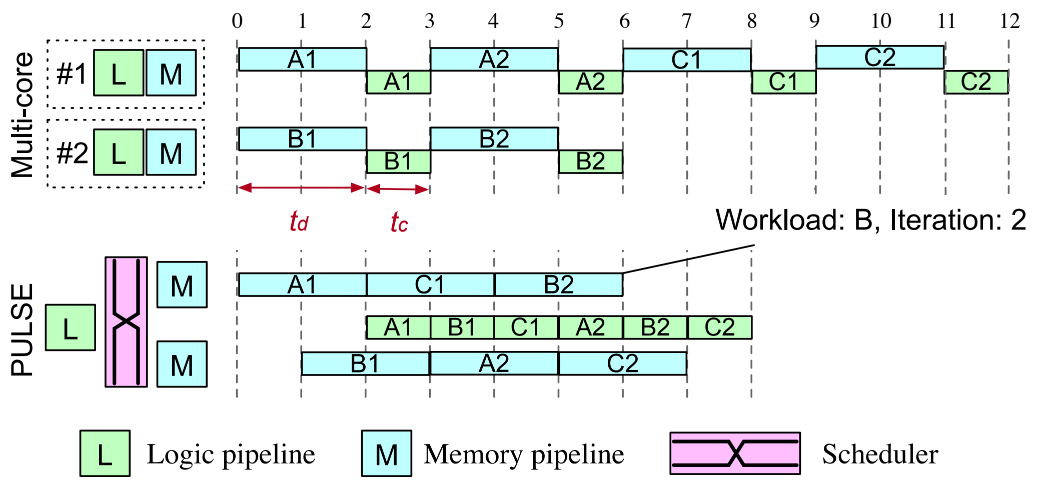


Fig. 3.5: **PULSE accelerator architecture.** (top) Traditional multi-core architectures with tightly coupled logic and memory pipelines result in low utilization and longer execution times. (bottom) PULSE accelerator's *disaggregated* design with an unequal number of logic and memory pipelines efficiently multiplexes concurrent iterator executions across them for near-optimal utilization and performance.

before the data fetch, and the next data fetch cannot be performed until the logic execution yields the next pointer.

- **Property 2:** Iterators that benefit from offload spend more time in data fetch (t_d) than logic execution (t_c), i.e., $t_c < \eta \cdot t_d$, where $\eta \leq 1$, as noted in §3.3.2.

Any accelerator designed for iterator executions must incorporate both a *memory pipeline* and a *logic pipeline* to support the execution steps (i) and (ii) mentioned earlier. However, the strict dependency between these two steps (Property 1) renders many traditional multi-core processor optimizations, such as out-of-order execution, ineffective. Additionally, because these architectures tightly couple logic and memory pipelines, the memory-intensive nature of iterators (Property 2) often results in the logic pipeline remaining idle for most of the time. These two factors together lead to poor resource utilization and energy inefficiency in such architectures.

Fig. 3.5 (top) illustrates this inefficiency using the execution of 3 iterators (A, B, C), each with 2 iterations (e.g., A1, A2, etc.), on a multi-core architecture. Since each iteration involves a data fetch followed by dependent logic execution, one pipeline (memory or logic) remains idle while the other is active. Although thread-level parallelism allows iterator requests to be distributed across multiple cores to increase overall throughput, the per-core underutilization of both logic and memory pipelines persists, resulting in suboptimal use of resources and increased energy consumption.

Disaggregated accelerator design. Motivated by the unique characteristics of iterators, we propose a novel accelerator architecture that *disaggregates memory and logic pipelines*, using a scheduler to multiplex iterator components across these pipelines. This decoupling enables an asymmetric number of logic and memory pipelines, maximizing the utilization of each, in contrast to the tightly coupled architecture of multi-core processors. In our design, if there are m logic pipelines and n memory pipelines, the accelerator-specific threshold $\eta < 1$ (as introduced in §3.3.2) is given by $\eta = \frac{m}{n}$, meaning there are fewer logic pipelines than memory pipelines, consistent with Property 2. Fig. 3.5 (bottom) illustrates an example of this disaggregated design with one logic pipeline and two memory pipelines ($m = 1, n = 2$).

Although data fetch and logic execution within each iterator must occur sequentially, the disaggregated architecture allows efficient multiplexing of data fetch and logic execution from different iterators across the separated pipelines, thus maximizing overall utilization. Recall that the logic execution time t_c for each offloaded iterator execution in PULSE is constrained by $t_c \leq \eta \cdot t_d$, where t_d is the time spent on data fetch (§3.3.2). In the extreme case where $t_c = \eta \cdot t_d$ for all iterator executions, it becomes possible to multiplex $m + n$ concurrent iterator executions to fully utilize all m logic and n memory pipelines. While we omit the theoretical proof for brevity, Fig. 3.5 (bottom) demonstrates the multiplexed execution—managed by a scheduler in our accelerator—for $t_c = \frac{1}{2} \cdot t_d$ using 3 iterators. This is the ideal case. Similar multiplexing is possible even when $t_c \leq \eta \cdot t_d$, fully utilizing the memory pipelines, though with lower utilization of logic pipelines (since they will be idle for a fraction of the time given by $\frac{t_c - \eta \cdot t_d}{t_c}$). Consequently, we provision $\eta = \frac{m}{n}$ to be as close as possible to the expected ratio $\frac{t_c}{t_d}$ for the workload to maximize the utilization of logic pipelines. Further improvements in logic pipeline energy efficiency can be achieved through dynamic frequency scaling [120], though we leave such optimizations for future work.

While the memory pipeline is stateless, the logic pipeline must maintain the state for the iterator it is executing. To efficiently multiplex several iterator executions, the logic pipelines require mechanisms for fast context switching. Each iterator execution is associated with a dedicated *workspace*, which stores three distinct pieces of state: `cur_ptr` and `scratch_pad`, which track the iterator state as described in §**??**, and `data`, which holds the memory data loaded for `cur_ptr`. Maintaining a dedicated workspace for each iterator allows the logic pipeline to switch between iterator executions without delay, as triggered by the scheduler. However, this requires maintaining multiple

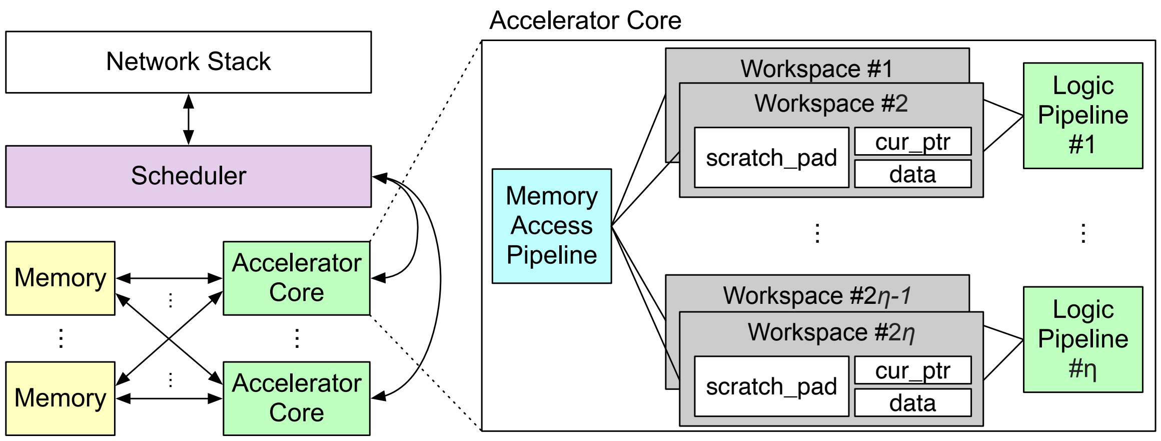


Fig. 3.6: **PULSE accelerator overview.** See §3.3.3 for details.

workspaces—up to $m + n$ to support all possible schedules, given our bound on the number of concurrent iterators. These workspaces are distributed evenly across the logic pipelines.

PULSE Accelerator Components. The PULSE accelerator consists of n memory pipelines and m logic pipelines for processing iterator requests, a scheduler that multiplexes these requests across the pipelines, and a network stack for handling pointer-traversal requests from the network (Fig. 3.6).

Memory pipeline: Each memory pipeline loads data from the attached DRAM to the corresponding workspace, as assigned by the scheduler at the start of each iteration. This process involves (i) address translation and (ii) memory protection based on page access permissions. To optimize on-chip storage, we implement range-based address translations, previously simulated in prior work [121], using TCAM.

Once a memory access is completed, the memory pipeline signals the scheduler to either continue the iterator execution or terminate it if a translation or protection failure occurs.

Logic pipeline: The logic pipeline executes all PULSE ISA instructions except for `LOAD`/`STORE`, determining the `cur_ptr` value for the next iteration or checking whether the termination condition has been met. Each logic pipeline consists of an ALU for executing arithmetic and logic instructions, along with modules for register manipulation, branching, and executing the specialized `RETURN` instruction (Table 3.2). During the execution of an iterator, the logic pipeline reads from and updates the registers in its dedicated workspace. An iteration can terminate in two ways: (i) the `cur_ptr` is updated to the next pointer and the `NEXT_ITER` instruction is reached, or (ii) the traversal completes and the `RETURN` instruction is reached. In either case, the logic pipeline sends the appropriate signal to the scheduler.

Scheduler: The scheduler coordinates the data fetch and logic execution for each iterator across the memory and logic pipelines:

1. Upon receiving a new request over the network, the scheduler assigns the iterator to an empty workspace in a logic pipeline and signals one of the memory pipelines to perform the data fetch from memory based on the state stored in the workspace.
2. After receiving a signal from the memory pipeline indicating that the data fetch has completed, the scheduler notifies the appropriate logic pipeline to continue the iterator execution using the corresponding workspace.
3. When the logic pipeline signals that the next iteration can begin (via the `NEXT_ITER` instruction), the scheduler signals one of the memory pipelines to execute the `LOAD` operation via the workspace.
4. If the scheduler receives a signal from the memory pipeline about an address translation or memory protection failure, or a signal from the logic pipeline indicating the iterator execution has reached its termination condition (via the `RETURN` instruction), it signals the network stack to prepare a response containing the iterator `code`, `cur_ptr`, and `scratch_pad`.

The scheduler assigns memory and logic pipelines in steps 1 and 3 to maximize the utilization of all memory pipelines (as illustrated in Fig. 3.5 (bottom)), though other scheduling policies could be implemented.

Network Stack: The network stack handles packet reception and transmission. When a new request arrives, it parses the payload to extract the request ID, `code`, and the state for the offloaded iterator execution (`cur_ptr`, `scratch_pad`).

The network stack uses the same format for both requests and responses, allowing it to send a response back to the CPU node upon traversal completion or to reroute the request to another memory node for further execution (§**??**).

Implementation. We implement PULSE on an FPGA-based NIC (Xilinx Alveo U250), which features two 100 Gbps Ethernet ports, 64 GB of on-board DRAM, 1,728K LUTs, and 70 MB of BRAM. The board's resources are partitioned into two PULSE accelerators, each utilizing one Ethernet port and two memory channels. Based on our analysis of common data structures (§3.4), which shows that the t_c/t_d ratio tends to be < 0.75, we configure $\eta = 0.75$. This results in four memory pipelines and three logic pipelines, with a total of 7 workspaces per accelerator.

For efficient operation, we use Xilinx TCAM IP [122] for page table management, 100 Gbps Ethernet IP, and link-layer IPs [123]. Additionally, burst data transfers [124] are employed to improve memory bandwidth. The logic and memory pipelines are clocked at 250 MHz, while the network stack operates at 322 MHz to handle 100 Gbps traffic. Although our current implementation demonstrates PULSE's capabilities on an FPGA prototype, we envision that the next logical step will be an ASIC implementation for even greater efficiency.

3.3.4 Distributed Pointer Traversals

Prior approaches to pointer traversals, which restrict them to a single memory node (§**??**), force applications into two less-than-ideal choices. On one hand, applications can confine their data to a single memory node, limiting scalability. On the other hand, they can distribute data across multiple nodes, but each time a pointer on a different memory node is accessed, the traversal must return to the CPU node. This latter approach improves scalability but introduces additional network and software processing latency at the CPU node.

To bypass the overhead of returning to the CPU node, one could replicate the entire translation and protection state across all memory nodes, allowing them to directly forward traversal requests to other memory nodes. However, this strategy comes with increased space requirements for storing the translation state, which is difficult to accommodate within the limited capacity of the accelerator's translation and protection tables. Moreover, replicating this state across memory nodes introduces complexity, requiring protocols to maintain consistency when state changes occur—protocols that impose significant performance overheads.

PULSE breaks the tradeoff between performance and scalability by utilizing a programmable network switch to enable rack-scale distributed pointer traversals. Specifically, when the PULSE accelerator at one memory node detects that the next pointer resides on a different memory node, it forwards the request to the network switch, which routes it to the correct memory node to continue the traversal. This approach reduces the network latency by half a round-trip time and eliminates the software overheads at the CPU node, as the routing logic is executed directly in the switch

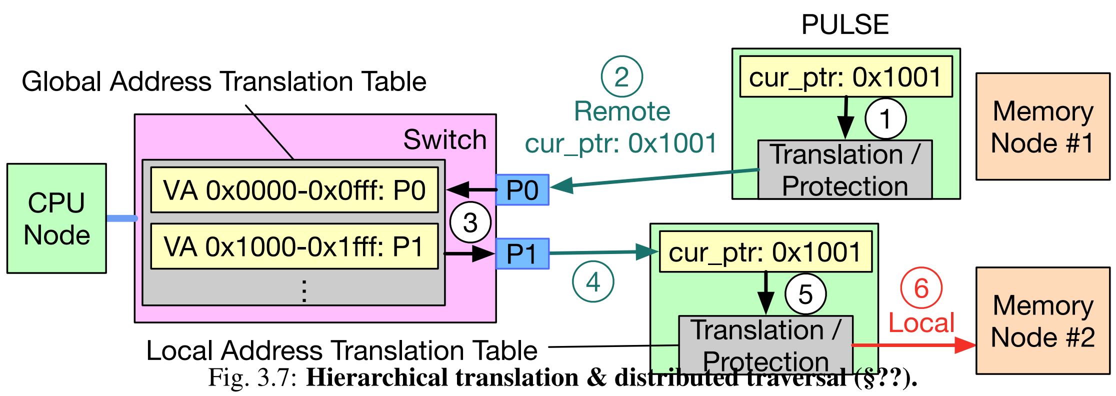


Fig. 3.7: **Hierarchical translation & distributed traversal (§??).**

hardware. Since routing traversal requests across memory nodes is analogous to packet routing, the switch hardware is already optimized for this process.

However, enabling rack-scale pointer traversals introduces two key challenges, which we address next.

Hierarchical translation. For the switch to forward a pointer traversal request to the correct memory node, it must determine which memory node is responsible for the relevant address. Given the limited resources at the switch, PULSE employs a hierarchical address translation mechanism, as illustrated in Fig. 3.7.

The address space is range-partitioned across memory nodes, and only the base address-to-memory node mapping is stored at the switch. Each memory node, in turn, maintains its local address translation and protection metadata at the accelerator (①), as described in §**??**. The switch inspects the `cur_ptr` field in the request (②) and uses its base address mapping to identify the target memory node (③). Once the request reaches the memory node, the traversal continues until a pointer is encountered that is not present in the local table (as in ①). At this point, the request is sent back to the switch (§3.3.3), which can either re-route the request to another memory node (④-⑥) or notify the CPU node if the pointer is invalid.

Continuing stateful iterator execution. A challenge in distributing iterator execution across memory nodes in PULSE is managing the stateful nature of the iterators. Since PULSE allows the storage of intermediate state in the iterator's `scratch_pad`, how can the execution of such stateful iterators be seamlessly continued on a different memory node? Fortunately, the design choice of confining all iterator state within the `scratch_pad` and `cur_ptr`, coupled with the use of consistent request and response formats, makes this straightforward. The accelerator at the current memory node embeds

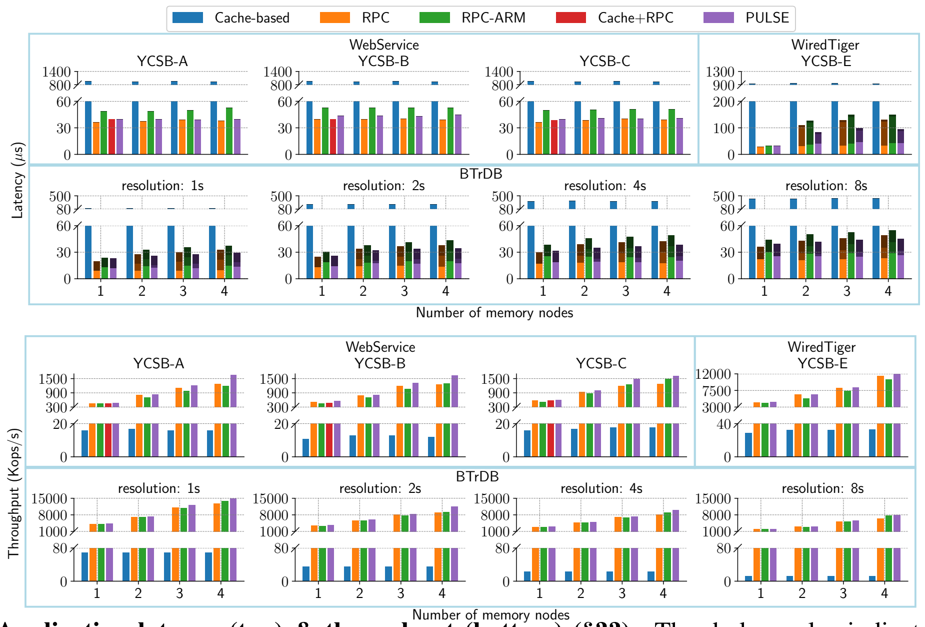


Fig. 3.8: **Application latency (top) & throughput (bottom) (§??).** The darker color indicates the time spent on cross-node pointer traversals, which increases with the number of memory nodes in WiredTiger and BTrDB.

the updated `scratch_pad` in the response before forwarding it to the switch. When the switch forwards the request to the next memory node, execution continues exactly as if the previous memory node had the pointer.

3.4 Evaluation

Compared systems. We compare PULSE against the following systems: (1) a **Cache-based** system that relies solely on CPU node caches to accelerate remote memory accesses, using Fastswap [31] as the representative system, (2) an **RPC** system that offloads pointer traversals to a CPU at the memory nodes, (3) **RPC-ARM**, an RPC system that employs wimpy ARM processors at the memory nodes, and (4) a **Cache+RPC** system that uses data structure-aware caches, represented by AIFM [28]. Systems (1) and (4) are configured with a cache size of 2 GB, while systems (2) and (3) use a DPDK-based RPC framework [125].

Our experimental setup consists of two servers—one acting as the CPU node and the other as the memory nodes—connected via a 32-port switch equipped with a 6.4 Tbps programmable Tofino ASIC. Both servers are powered by Intel Xeon Gold 6240 processors [126] and feature 100 Gbps Mellanox ConnectX-5 NICs.

Application	Data Structure	t_c/t_d	#Iterations
WebService	Hash-table	0.06	48
WiredTiger	B+Tree	0.63	25
BTrDB ($1s$ to $8s$)		0.71	38–227

Table 3.3: **Workloads used in our evaluation (§3.4).** t_c and t_d correspond to compute and memory access time at the PULSE accelerator.

To ensure a fair comparison, we limit the memory bandwidth of the memory nodes to 25 GB/s, which corresponds to the peak bandwidth of the FPGA, using Intel Resource Director Technology [127]. We report the energy consumption based on the **minimum** number of CPU cores required to saturate the available bandwidth. For the ARM-based system (**RPC-ARM**), we use Bluefield-2 DPUs [128], which feature 8 Cortex-A72 cores and 16 GB of DRAM.

For PULSE, we configure two memory nodes per FPGA NIC (one per port), resulting in a total of four memory nodes. The results from our experiments can be extrapolated to larger setups, as PULSE's performance and energy efficiency are independent of dataset size and cluster scale.

Applications & workloads. We consider 3 applications with varying data structure complexity, compute/memory-access ratio, and iteration count per request (Table 3.3): (1) *Web Service* [28] that processes user requests by retrieving user IDs from an in-memory hash table, using these IDs to fetch 8KB objects, which are then encrypted, compressed and returned to the user. Requests are generated using YCSB A (50% read/50% update), B (95% read/5% update), and C (100% read) workloads with Zipf distribution [129]. (2) *WiredTiger Storage Engine* (MongoDB backend [130]) uses B+Trees to index NoSQL tables. Our frontend issues range query requests over the network to WiredTiger and plots the results. Similar to prior work [28, 131], we model user queries using the YCSB E workload with Zipf distribution [129] on 8B keys and 240B values. (3) *BTrDB Time-series Database* [108] is a database designed for visualizing patterns in time-series data. BTrDB reads the data from a B+Tree-based store for a given user query and renders the time-series data through an interactive user interface [132]. We run stateful aggregations (sum, average, min, max) for time windows of different resolutions, from 1s to 8s, on the Open μPMU Dataset [133] with voltage, current, and phase readings from LBNL's power grid [108].

Performance for Real-world Applications

Since AIFM [28] does not natively support B+-Trees or distributed execution, we restrict the Cache+RPC approach to the Web Service application on a single node.

Single-node performance. Fig. 3.8 demonstrates the advantages of accelerating pointer-traversals at disaggregated memory. Compared to the Cache-based approach, PULSE achieves 9–34.4× lower latency and 28–171× higher throughput across all applications using only one network round-trip per request. RPC-based systems observe 1–1.4× lower latency than PULSE due to their 9× higher CPU clock rates. We believe an ASIC-based realization of PULSE has the potential to close or even overcome this gap. Cache+RPC incurs higher latency than RPC due to its TCP-based DPDK stack [28, 134] and does not outperform RPC, indicating that data structure-aware caching is not beneficial due to poor locality.

Latency depends on the number of nodes traversed during a single request and the response size. WebService experiences the highest latency due to large 8KB responses and long traversal length per request. In BTrDB, the latency increases (and the throughput decreases) as the window size grows due to the longer pointer traversals (see Table 3.3). Interestingly, the Cache-based approach performs significantly better for BTrDB than WebService and WiredTiger due to the better data locality in time-series analysis of chronologically ordered data. However, its throughput remains significantly lower than both PULSE and RPC since it is bottlenecked by the swap system performance, which could not evict pages fast enough to bring in new data. This is verified in our analysis of resource utilization (deferred to Appendix for brevity); we find that RPC, RPC-ARM, Cache+RPC, and PULSE can utilize more than 90% of the memory bandwidth across the applications, while the Cache-based approach observes less than 1 Gbps network bandwidth. The other systems — PULSE, RPC, RPC-ARM, and Cache+RPC — can also saturate available memory bandwidth (around 25 GB/s) by offloading pointer traversals to the memory node, consuming only 0.5%–25% of the available network bandwidth.

Distributed pointer traversals. Fig. 3.8 shows that employing multiple memory nodes introduces two major changes in performance trends: (1) the latency increases when the pointer traversal spans multiple memory nodes, and (2) throughput increases with the number of nodes since the systems can exploit more CPUs or accelerators. WebService is an exception to the trend: since the hash

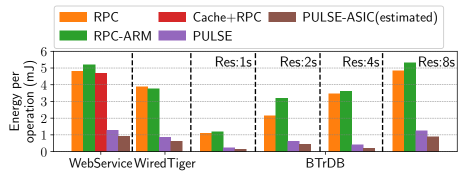


Fig. 3.9: **Application energy consumption per operation (§??).**

table is partitioned across memory nodes based on primary keys, the linked list for a hash bucket resides in a single memory node.

PULSE observes lower latency than the compared systems due to in-network support for distributed pointer-traversals (§**??**). The latency increases significantly from one to two memory nodes for all systems since traversing to the next pointer on a different memory node adds 5–10 μs network latency. Also, even across two memory nodes, a request can trigger multiple inter-node pointer traversals incurring multiple network round-trips; for WiredTiger and BtrDB, 10%–30% of pointer traversals are inter-node. However, in-network traversals allow PULSE to reduce latency overheads by 33–98%, with 1.1–1.36$\times$ higher throughput than RPC.

Energy consumption. We compared energy consumed per request for PULSE and RPC schemes at a request rate that ensured memory bandwidth was saturated for both. We measure energy consumption using Xilinx XRT [135] for PULSE (all power rails) and Intel RAPL tools [136] for RPC on CPUs [126] (CPU package and DRAM only). For RPC-ARM on ARM cores, since there is no power-related performance counter [137] or open-source tool available, we adapt the measurement approach from prior work [138]. Specifically, we calculate the CPU package's energy using application CPU cycle counts and DRAM power using Micron's estimation tool [139]. Finally, we conservatively estimate ASIC power using our FPGA prototype: we scale down the ASIC energy only for PULSE accelerator using the methodology employed in prior research [140] while using the unscaled FPGA energy for other components (DRAM, third-party IPs, etc.). As such, we measure an *upper bound* on PULSE and PULSE-ASIC energy use, and a *lower bound* for RPC, RPC-ARM, and Cache+RPC.

Fig. 3.9 shows that PULSE achieves a 4.5–5$\times$ reduction in energy use per operation compared to RPCs on a general-purpose CPU, due to its disaggregated architecture (§3.3.3). Our estimation shows that PULSE's ASIC realization can conservatively reduce energy use by an additional 6.3 – 7$\times$ factor. Finally, RPC-ARM's total energy consumption per request can exceed that of standard

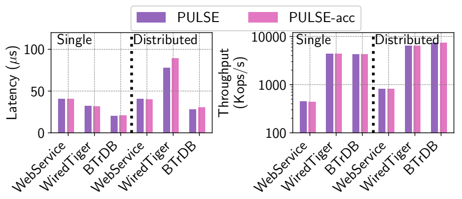


Fig. 3.10: **Impact of distributed pointer traversals (§??).**

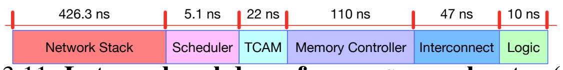


Fig. 3.11: **Latency breakdown for PULSE accelerator (§??).**

cores, as seen in the WebService workload. This observation aligns with prior studies [138], which attribute the increased energy use to their longer execution times, resulting in higher aggregate energy demands.

Understanding PULSE Performance

Distributed pointer traversals. We evaluate the impact of distributed pointer traversals (§**??**) by comparing PULSE against PULSE-ACC, a PULSE variant that sends requests back to the CPU node if the next pointer is not found on the memory node. Fig. 3.10 shows that while both have identical performance on a single memory node, PULSE-ACC observes 1.02–1.15× higher latency for two nodes. On the other hand, their throughput is the same since, under sufficient load, memory node bandwidth bottlenecks the system for both.

Latency breakdown for PULSE accelerator. Fig. 3.11 shows the latency contributions of various hardware components at the PULSE accelerator for the WebService application. The network stack first processes the pointer traversal request in about 430 ns, after which the WebService payload is processed by the scheduler and dispatched to an idle memory access pipeline in 5.1 ns. Then, the memory pipeline takes $\sim$132 ns to perform address translation, memory protection, and data fetch from DRAM. Finally, the logic pipeline takes 10 ns to check the termination conditions and determine the next pointer to look up. This process repeats until the termination condition is met. The time to send a response back over the network stack is symmetric to the request path.

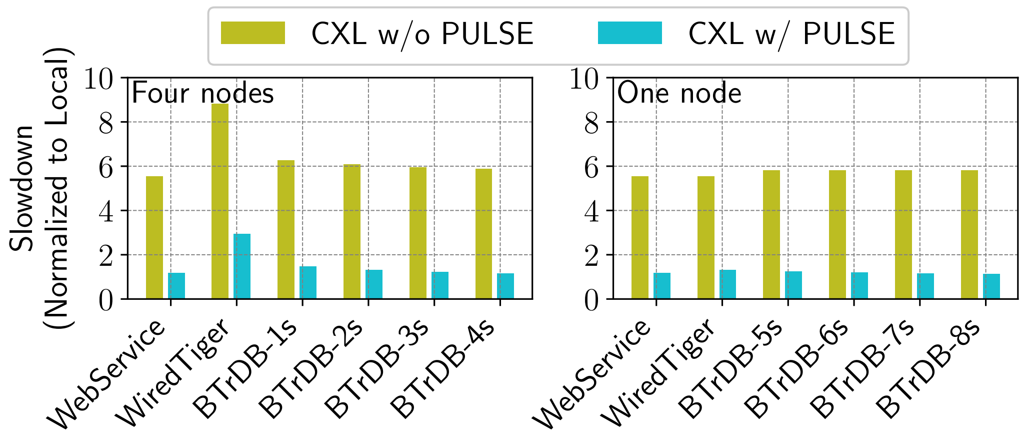


Fig. 3.12: **Slowdown with simulated CXL interconnect (§3.6).**

3.5 Related Work

Prior work has explored such processing units in near-memory and processing-in-memory architectures [141–169], as well as CPUs [28, 170–174] and FPGAs [138, 175] near remote/disaggregated memory, though these approaches have notable limitations.

Shortcomings of Prior Approaches.

No prior work achieves all four properties required for pointer traversals on disaggregated memory: distributed execution, expressiveness, energy efficiency, and performance. We focus on network-attached memory, although a similar analysis extends to in-memory processing [141–169].

No support for distributed execution. Distributed pointer traversals are required to ensure applications can efficiently access large pools of network-attached memory nodes. Unfortunately, to our knowledge, none of the prior works support efficient multi-node pointer traversals. Therefore, applications must confine their data to a single node for efficient traversals, exposing a tradeoff between application performance and scalability. Recent proposals [29,109,176–180] explore specialized data structures that co-design partitioning and allocation policies to reduce distributed pointer traversals atop disaggregated memory. Such approaches complement our work since they still require efficient distributed traversals when their optimizations are not applicable, *e.g.*, not many data structures benefit from such specialized co-designs.

Poor utilization/power-efficiency in CPUs. Many prior works have explored remote procedure call (RPC) interfaces to enable offloading computation to CPUs on memory nodes [28, 170–173].

While CPUs are performant and versatile enough to support most general-purpose computations, the same versatility makes them overkill for pointer traversal workloads in disaggregated architectures — the CPUs on memory nodes are likely to be underutilized and, consequently, waste energy (§3.4), since such workloads are memory-intensive and bounded by memory bandwidth rather than CPU cycles. Since inefficient power usage resulting from coupled compute and memory resources is the main problem disaggregation aims to resolve, leveraging CPUs at memory nodes essentially nullifies these benefits.

Limited expressiveness in FPGA/ASIC accelerators. Another approach explored in recent years uses FPGAs [138, 175] or ASICs [168, 169] at memory nodes for performance and energy efficiency. FPGA approaches exploit circuit programmability to realize performant on-path data processing, albeit only for specific data structures, limiting their expressiveness. Although some FPGA approaches aim for greater expressiveness by serving RPCs [181], RPC logic must be pre-compiled before it is deployed and physically consumes FPGA resources. This limits how many RPCs can be deployed on the FPGA concurrently and also elides runtime resource elasticity for different pointer traversal workloads. ASIC approaches either support a single data structure or provide limited ISA specialized for a single data structure (*e.g.*, linked-lists [168]), limiting their general applicability.

Poor performance/power efficiency in wimpy SmartNICs. The emergence of programmable SmartNICs has driven work on offloading computations to the onboard network processors. Some approaches utilize wimpy processors (*e.g.*, ARM or RISC-V processors) [182] or RDMA processing units (PUs) [183] to support general-purpose computations near memory. While these wimpy processors can eliminate multiple network round trips in pointer traversal workloads, their processing speeds are far slower than CPU-based or FPGA-based accelerators. Often, such PUs can become a performance bottleneck, especially at high memory bandwidth ($\sim$500 Gbps) [3, 183]. Moreover, wimpy processors tend not to be energy-efficient since their slower execution tends to waste more static power, resulting in higher energy per pointer traversal offload — an observation noted in prior work [138] and confirmed in our evaluation (§3.4).

Specifically, existing approaches are limited in scale and expose a three-way tradeoff between expressiveness, energy efficiency, and performance. First, and perhaps most crucially, none of the existing approaches can accelerate pointer traversals that span *multiple* network-attached memory nodes.

This limits memory utilization and elasticity since applications must confine their data to a single memory node to accelerate pointer traversals. Their inability to support distributed pointer traversals stems from complex management of address translation state that is required to identify if a traversal can occur locally or must be re-routed to a different memory node (§**??**). Second, existing single-node approaches use full-fledged CPUs for expressive and performant execution of pointer-traversals [28, 170–172]. However, coupling large amounts of processing capacity with memory — which has utility in reducing data movement in PIM architectures [141–153] — goes against the very spirit of memory disaggregation since it leads to poor utilization of compute resources and, consequently, poor energy efficiency.

Approaches that use wimpy processors at SmartNICs [182, 183] instead of CPUs retain expressiveness, but the limited processing speeds of wimpy nodes curtail their performance and, ultimately lead to lower energy efficiency due to their lengthened executions (§**??**, [138]). Lastly, FPGA-based [138, 175, 184] and ASIC-based [168, 169] approaches achieve performance and energy efficiency by hard-wiring pointer traversal logic for specific data structures, limiting their expressiveness.

3.6 Summary

While PULSE is implemented atop Ethernet, its design is interconnect-agnostic and could be realized in ASIC-based or FPGA-attached memory devices over emerging interconnects like CXL [10, 184, 185]. We have verified these benefits in simulation atop detailed memory access and processing traces of our evaluated applications and workloads. The simulator maintains 2GB of cache in local (CPU-attached) DRAM, while the entire working set is stored on remote CXL memory. Following prior work [37], we model 10–20ns L3 cache latency, 80ns local DRAM latency, 300ns CXL-attached memory latency, and 256B access granularity. We simulate both a four-memory-node setup, which uses a CXL switch with PULSE logic and a PULSE accelerator at each memory node, and a single-node setup with no switch. We assume a conservative overhead for PULSE, using our hardware programmable Ethernet switch and FPGA accelerator latencies.

Fig. 3.12 shows the average slowdown for executing our evaluated workloads on CXL memory relative to running it completely locally (i.e., the entire application working set fits in local DRAM) — with and without PULSE. In the four-node setup, PULSE reduces CXL's slowdown by 19–33% across all applications.

In the single-node setup, PULSE still reduces the slowdown by 19–23% by minimizing high-latency traversals over the CXL interconnect. While a real hardware realization is necessary to precisely quantify PULSE's benefits, our simulation (which models the lowest possible CXL latency and highest possible PULSE overheads) highlights its potential for improving performance in emerging interconnects.

Chapter 4

Hardware Layer: Memory Management for Next-Gen Interconnects

While MIND and PULSE implemented memory management functionality over Ethernet, and network-based resource disaggregation has gained traction due to advancements in network bandwidth, the inherent latency—limited by the speed of light—continues to impose significant overhead. Recent hardware advancements have led to the development of new-generation interconnects by major hardware vendors, such as Nvidia's NVLink [186] and Intel's Compute Express Link (CXL) [10]. CXL, in particular, has emerged as a promising solution for expanding memory capacity and bandwidth by attaching external memory devices to PCIe slots, offering a dynamic and heterogeneous computing environment. Its low-latency and scalable nature make CXL an ideal interconnect for disaggregated architectures.

There are several fundamental differences between CXL and Ethernet, which we summarize below:

1. **Data transfer mechanism**: Ethernet uses packet-based transmission, where data is encapsulated in frames with headers and footers, potentially increasing latency due to overhead. CXL, in contrast, provides memory semantics, enabling faster and more efficient data transfers without the overhead associated with packet framing.
2. **Performance**: CXL offers orders of magnitude faster performance and provides significantly higher memory bandwidth compared to Ethernet.

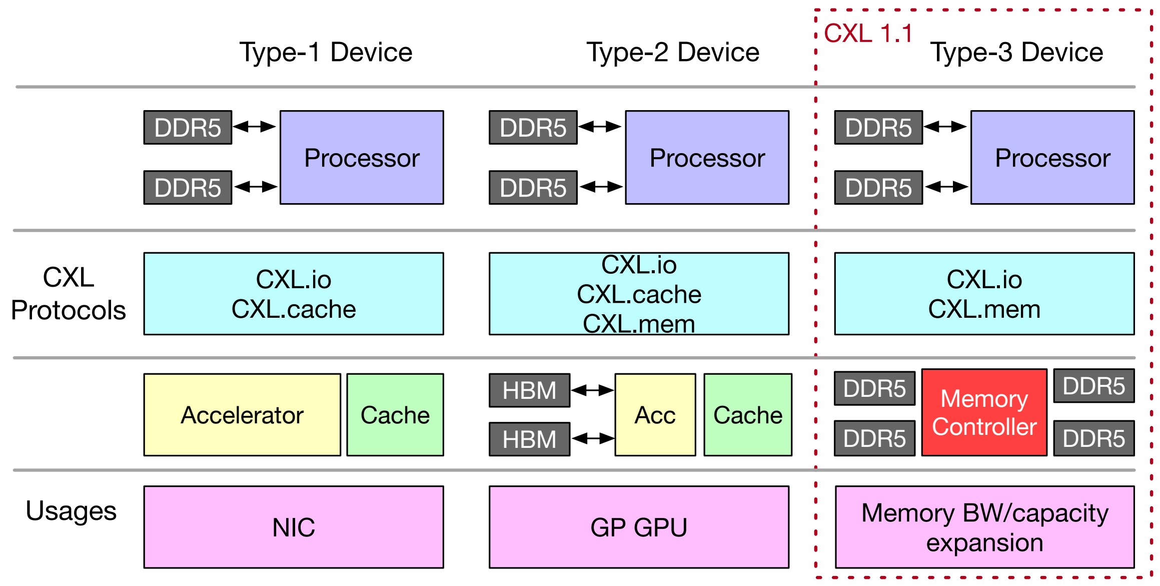


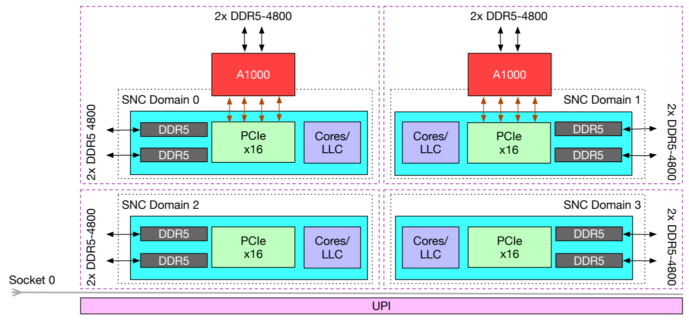
Fig. 4.1: **CXL Overview.** In this study, we focus on commercial CXL 1.1 Type-3 devices, leveraging CXL.io and CXL.mem protocols for memory expansion in single-server environments.

3. **Scale**: Current CXL prototypes [187, 188] are limited to operating within a single rack, whereas Ethernet can scale across entire data centers.

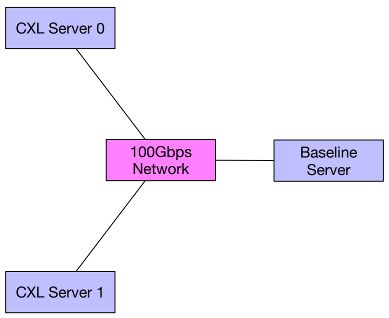
Given these fundamental differences between the two interconnects, directly applying MIND's or PULSE's techniques to a CXL-based disaggregated architecture presents challenges. For instance, it is unclear whether an RMT-style packet switching network could be implemented within a CXL switch. This chapter explores the potential of next-generation interconnects like CXL and examines how the software stack must adapt to leverage these new technologies.

We begin by presenting an empirical study of the latest CXL ASIC prototypes and investigate their potential application in modern data center environments. We then discuss ongoing efforts to deploy CXL in real-world infrastructure and applications.

Our study aims to fill existing knowledge gaps by conducting detailed evaluations of CXL 1.1 for memory-intensive applications, leading to several *intriguing observations*: Contrary to the common perception that CXL memory, due to its higher latency, should be considered a separate, slower tier of memory [37, 189], **we find that shifting some workloads to CXL memory can significantly enhance performance**, even if local memory's capacity and bandwidth are underutilized. This is because using CXL memory can decrease the overall memory access latency by alleviating bandwidth contention on DDR channels, thereby improving application performance. From our analysis



(a) CXL Server(socket 0 illustrated here).



(b) Server setup.

Fig. 4.2: **CXL Experimental Platform.** (a) Each CXL server is equipped with two A1000 memory expansion cards. SNC-4(§4.2.1) is enabled only for the raw performance benchmarks(§4.2) and bandwidth-bound benchmarks(§4.4), and each SNC Domain is equipped with two DDR5 channels. (a) illustrates Socket 0; Socket 1 shares a similar setup except for the absence of CXL memory. (b) Our platform comprises two CXL servers and one baseline server. The baseline server replicates the same configuration but lacks any CXL memory cards.

of application performance, we have formulated an abstract cost model (§4.5) that predicts substantial cost savings in practical deployments.

4.1 Background and Methodology

This section presents an overview of CXL technology, followed by our experimental setup and methodologies.

Compute Express Link (CXL) is a standardized interconnect technology that enables communication between processors and various devices, such as accelerators, memory expansion units, and smart I/O devices. CXL is built upon the physical layer of PCI Express® (PCIe®) 5.0 [190], offering native support for x16, x8, and x4 link widths with data rates of 32.0 GT/s and 64.0 GT/s. The CXL transaction layer is implemented through three protocols: CXL.io, CXL.cache, and CXL.mem, as shown in Fig. 4.1. The *CXL.io* protocol, based on PCIe 5.0, handles device discovery, configuration, initialization, I/O virtualization, and direct memory access (DMA). *CXL.cache* enables CXL devices to access the host processor's memory, while *CXL.mem* allows the host to access device-attached memory using load/store commands.

CXL devices are classified into three types, each suited to specific use cases:

1. *Type-1 devices*, such as SmartNICs, utilize CXL.io and CXL.cache for communication with DDR memory.

2. *Type-2 devices*, including GPUs, ASICs, and FPGAs, use CXL.io, CXL.cache, and CXL.mem to share memory with the processor, enhancing workloads within the same cache domain.

3. *Type-3 devices* leverage CXL.io and CXL.mem for memory expansion and pooling, allowing for increased DRAM capacity, enhanced memory bandwidth, and the addition of persistent memory without occupying DRAM slots. These devices augment DRAM with CXL-enabled solutions, providing high-speed, low-latency storage.

The current commercially available version of CXL is 1.1, which limits each CXL 1.1 device to function as a single logical device accessible by only one host at a time. Future generations, such as CXL 2.0, are expected to support partitioning devices into multiple logical units, allowing up to 16 different hosts to access separate portions of memory [191]. In this work, we focus on commercially available CXL 1.1 Type-3 devices, specifically addressing their use for single-host memory expansion.

4.1.1 Hardware Support for CXL

Recent announcements have introduced CXL 1.1 support for Intel Sapphire Rapids processors (SPR) [192] and AMD Zen 4 EPYC "Genoa" and "Bergamo" processors [193]. While commercial CXL memory modules are available from vendors such as Asteralabs [194], Montage [195], Micron [139], and Samsung [196], CXL memory expanders are still primarily in the prototype stage, with limited samples available, making access difficult for university research labs. As a result, due to the scarcity of CXL hardware, much of the research into CXL memory has relied on NUMA-based emulation [37, 189] and FPGA implementations [187, 197], each presenting certain limitations:

NUMA-based emulation. Given the cache-coherent nature and comparable transfer speeds between CXL and UPI/xGMI interconnects, NUMA-based emulation [37,189] has been widely adopted for rapid application performance analysis and software prototyping. In this approach, CXL memory is exposed as a remote NUMA node. However, NUMA-based emulation fails to capture the precise performance characteristics of CXL memory due to inherent differences between CXL and UPI/xGMI interconnects [198], as highlighted in previous research [187].

FPGA-based implementation. Some hardware vendors, including Intel, use FPGA hardware to implement CXL protocols [199], overcoming the performance inconsistencies of NUMA-based emulation. However, FPGA-based CXL memory implementations do not fully exploit memory chip performance due to the lower operating frequencies of FPGAs compared to ASICs [200]. While FPGAs offer flexibility, they prioritize versatility over performance, making them suitable for early-stage CXL memory validation but not for production deployment. Intel's recent evaluation [187] revealed several performance limitations in FPGA-based implementations, including reduced memory bandwidth during concurrent thread execution. This hampers rigorous evaluations for memory capacity- and bandwidth-bound applications, which are critical use cases for CXL memory expanders. A detailed discussion on the performance gap between CXL ASIC and FPGA controllers is provided in §4.2.

4.1.2 Software Support for CXL

While hardware vendors are actively advancing CXL production, a notable gap remains in software and OS kernel support for CXL memory. This deficiency has driven the development of specific software enhancements. We summarize the most recent patches in the Linux Kernel that add CXL-aware support, namely: (1) the interleaving policy support (unofficial) and (2) the hot page selection support (official since Linux Kernel v6.1).

N:M Interleave Policy for Tiered Memory Nodes.

Traditional memory interleave policies distribute data evenly across memory banks, typically using a 1:1 ratio. However, with the emergence of tiered memory systems—where CPU-less memory nodes exhibit varying performance characteristics—new strategies are required to optimize memory bandwidth for bandwidth-intensive applications. The interleave patch [201] introduces an N:M interleave policy, which allows for the allocation of N pages to high-performance (top-tier) memory nodes and M pages to lower-tier nodes. For example, a 4:1 ratio directs 80% of traffic to top-tier nodes and 20% to lower-tier nodes. This ratio can be adjusted using the `vm.numa_tier_interleave` parameter. While the patch shows promising evaluation results [201], the optimal memory distribution depends on specific hardware and application characteristics. Given the higher latency of CXL memory, as demonstrated in §4.2, performance-sensitive applications must be carefully profiled and benchmarked to fully leverage interleaving while mitigating potential performance trade-offs.

NUMA Balancing & Hot Page Selection.

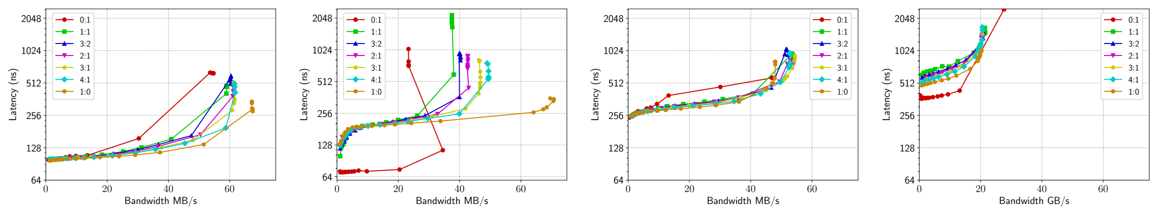
The memory subsystem, now termed a memory tiering system, accommodates various memory types like PMEM and CXL memory, each with differing performance characteristics. To optimize system performance, frequently accessed "hot pages" should reside in faster memory tiers like DRAM, while less frequently accessed "cold pages" should be placed in slower tiers like CXL memory. Recent Linux Kernel patches address this:

1. The *NUMA-balancing* patch [202] implements a latency-aware page migration strategy that promotes recently accessed (MRU) pages by scanning NUMA balancing page tables and hinting at page faults. However, it may fail to accurately identify high-demand pages due to long scanning intervals, which could cause latency issues for certain workloads.
2. The *Hot Page Selection* patch [203] introduces a Page Promotion Rate Limit (PPRL) mechanism to control the rate at which pages are promoted or demoted. While this extends the time for promotions/demotions, it improves workload latency by dynamically adjusting the hot page threshold to align with the promotion rate limit.

Additionally, research prototypes like TPP [189] employ similar optimization concepts and are being considered for integration into the Linux Kernel [204]. However, during our testing with memory bandwidth-intensive applications, we encountered unexplained performance degradation with TPP. As a result, we rely on the well-tested kernel patches integrated into Linux Kernel since version 6.1.

4.1.3 Experimental Platform Description

Our evaluation testbed, illustrated in Fig. 4.2(b), consists of three servers. Two of these servers are dedicated to CXL experiments and are equipped with dual Intel Xeon 4th Generation CPUs (Sapphire Rapids, SPR), 1 TB of 4800 MHz DDR5 memory, two 1.92 TB SSDs, and two A1000 CXL Gen5 x16 ASIC memory expander modules from AsteraLabs, each equipped with 256 GB of 4800 MHz memory (for a total of 512 GB of memory per server). Both A1000 modules are attached to socket 0.



(a) Local-socket MMEM (b) Remote-socket MMEM (c) Local-socket CXL (d) Remote-socket CXL

Fig. 4.3: **Overall effect of read-write ratio on MMEM and CXL across different distances.** The workloads are represented by read:write ratios (e.g., 0:1 for write-only, 1:0 for read-only). Accessing CXL memory locally incurs higher latency compared to MMEM but is more comparable to accessing MMEM on a remote socket. MMEM bandwidth peaks at 67 GB/s, versus 54.6 GB/s for CXL memory. Performance significantly declines when accessing CXL memory on a remote socket (§4.2.2). In specific scenarios, such as the write-only workload (0:1) in (b), the plot may show instances where bandwidth decreases and latency increases with heavier loads. The Y-axis is on a logarithmic scale.

The third server serves as the baseline and is configured identically to the CXL experiment servers, except it lacks CXL memory expanders. It is used to initiate client requests and run workloads that strictly utilize main memory during application assessments. All servers are interconnected via 100 Gbps Ethernet links.

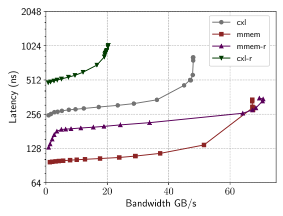
4.2 CXL 1.1 Performance Characteristics

In this section, we assess the performance of the CXL memory expander and compare it directly with main memory, which we designate as **MMEM** for clarity when contrasted against CXL memory. We analyze workload patterns and evaluate performance differences between local and remote socket scenarios.

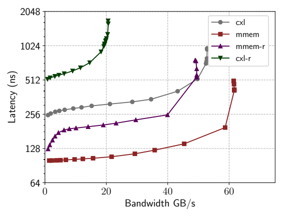
4.2.1 Experimental Configuration

For each dual-channel A1000 ASIC CXL memory expander [194], we connect two DDR5-4800 memory channels, providing a total capacity of 256 GB. To ensure a fair comparison between MMEM and CXL-attached DDR5 memory, we use the Sub-NUMA Clustering (SNC) [205] feature to equalize the number of memory channels in both configurations.

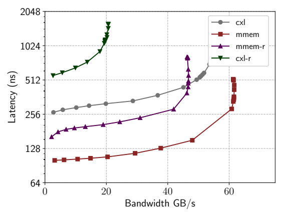
Sub-NUMA Clustering (SNC). Sub-NUMA Clustering (SNC) is an enhancement over the traditional NUMA architecture, dividing a single NUMA node into smaller semi-independent sub-nodes (domains). Each sub-NUMA node has its own dedicated local memory, L3 caches, and CPU cores.



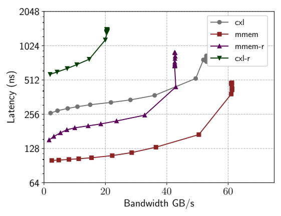
(a) Read-only workload



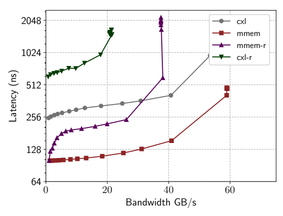
(b) Read:Write = 4:1 workload



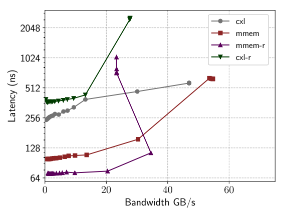
(c) Read:Write = 3:1 workload



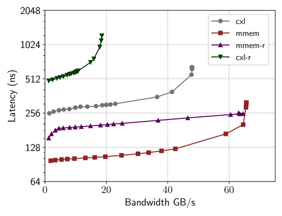
(d) Read:Write = 2:1 workload



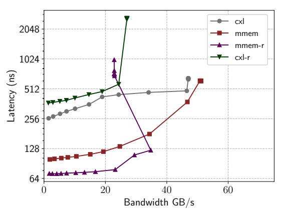
(e) Read:Write = 1:1 workload



(f) Write-only workload



(g) Read-only workload (random)



(h) Write-only workload (random)

Fig. 4.4: **A detailed comparison of MMEM versus CXL over diverse NUMA/socket distances and workloads.** (a)-(f) shows the latency-bandwidth trend difference of accessing data from different distances in sequential access pattern, sorted by the proportion of write. We refer to main memory as **MMEM**, with MMEM-r and CXL-r representing remote socket MMEM and cxl memory access, respectively. The Y-axis is on a logarithmic scale.

In our experimental setup (Fig. 4.2(a)), each CPU is partitioned into four sub-NUMA nodes. Each sub-NUMA node is equipped with two DDR5 memory channels connected to two 64 GB DDR5-4800 DIMMs. SNC is enabled by setting the Integrated Memory Controllers (IMC) to 1-way interleaving. Based on specifications, a single DDR5-4800 channel has a theoretical peak bandwidth of 38.4 GB/s [206], resulting in a combined memory bandwidth of up to 76.8 GB/s per sub-NUMA node.

Intel Memory Latency Checker (MLC). We use Intel's Memory Latency Checker (MLC) to measure loaded-latency for various read-write workloads, employing a 64-byte access size, consistent with prior work [187]. We deploy 16 MLC threads, and while the thread count in MLC is configurable, it does not directly control memory request concurrency. Instead, MLC assigns distinct memory segments to each thread for simultaneous access. When evaluating loaded latency, MLC incrementally increases the operation rate per thread. Our results show that using 16 threads with MLC accurately measures both idle and loaded latency, as well as the point at which bandwidth saturation occurs. MLC supports a wide range of workloads, including different read-write mixes and non-temporal writes.

Our study aims to address the following research questions:

- How does the performance of CXL-attached memory compare to local-socket and remote-socket main memory?
- What is the performance impact of CXL memory under different read-write ratios and access patterns (random vs. sequential)?
- How do main memory and CXL memory behave under high memory load conditions?

4.2.2 Basic Latency and Bandwidth Characteristics

This section presents our findings on memory access latency and bandwidth across different memory configurations: local-socket main memory (MMEM), remote-socket main memory (MMEM-r), CXL memory (CXL), and remote-socket CXL memory (CXL-r). Figure 4.3(a) illustrates the loaded latency curve for MMEM under various read-write mixes. The read-only workload achieves a peak bandwidth of approximately 67 GB/s, reaching 87% of its theoretical maximum. However, as the proportion of write operations increases, bandwidth decreases, with write-only workloads dropping to 54.6 GB/s. Initial memory latency is about 97 ns, but it rises sharply as bandwidth nears full capacity, indicating bandwidth contention [207, 208]. Interestingly, latency starts increasing significantly at 75%-83% of bandwidth utilization, exceeding prior estimates of 60% from earlier studies [207].

Figure 4.3(b) compares latency for MMEM when accessed via a remote socket. For read-only workloads, latency starts around 130 ns, whereas write-only workloads exhibit a much lower latency of 71.77 ns. This reduction in latency for write-only operations stems from non-temporal writes, which proceed asynchronously without waiting for confirmation. While read-only tasks achieve similar maximum bandwidth to local MMEM, increasing the proportion of write operations drastically reduces bandwidth due to additional UPI traffic generated by cache coherence protocols. Notably, write-only workloads generate minimal UPI traffic but suffer from the lowest bandwidth because they utilize only one direction of the UPI's bidirectional capacity. Moreover, latency escalation occurs earlier in remote-socket memory accesses than in local-socket ones, primarily due to queue contention at the memory controller.

Figure 4.3(c) illustrates the latency curve for CXL memory expansion, showing a minimum latency of 250.42 ns. Despite the added overhead from PCIe and the CXL memory controller, CXL follows a similar "bandwidth contention" pattern as MMEM. Latency remains relatively stable as bandwidth increases, with a maximum of 56.7 GB/s achieved under a 2 : 1 read-write ratio. The lower maximum bandwidth compared to DRAM is attributed to PCIe overhead, such as additional headers. For read-only workloads, the maximum bandwidth is further reduced due to PCIe's bidirectional nature, preventing full bandwidth utilization. Figure 4.3(d) displays the latency-bandwidth relationship for remote-socket CXL access, revealing an idle latency as high as 485 ns. Additionally, maximum memory bandwidth is unexpectedly halved, reaching only 20.4 GB/s under a 2 : 1 read-write ratio— a much more severe performance drop compared to remote-socket MMEM (Fig. 4.3(d)). Since read-only access to a CXL Type-3 device on a remote socket does not generate significant coherence traffic, cache coherence can be ruled out as a cause. Further investigation using Intel Performance Counter Monitor (PCM) [209] confirmed that UPI utilization remained consistently below 30%. Discussions with Intel suggest this bottleneck likely stems from limitations in the Remote Snoop Filter (RSF) on the current CPU platform, which may be addressed in next-generation processors [210].

4.2.3 Different Read-Write Ratios & Access Patterns

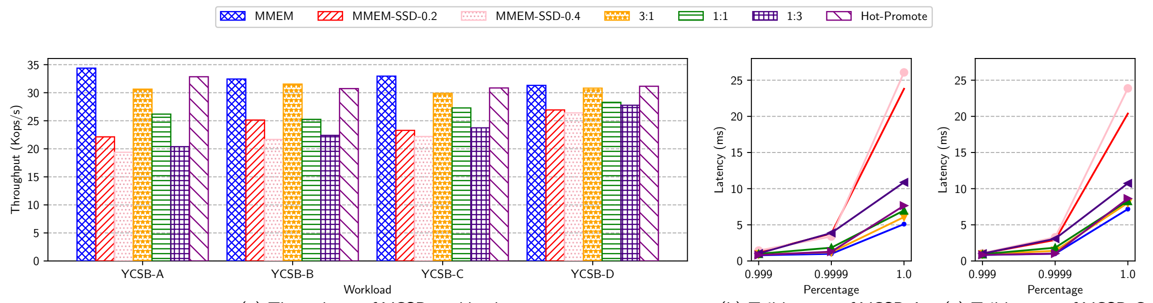
Figures 4.4(a)–4.4(f) compare performance under various read-write ratios. The results support our earlier observation that accessing CXL from a remote socket results in significantly higher latency and lower bandwidth. When accessing CXL from the same socket, latency is 2.4-2.6 times that of local DDR and 1.5-1.92 times that of remote-socket DDR. This suggests that directly running applications on CXL memory could severely degrade performance. However, for workloads spanning multiple NUMA nodes within the same socket, accessing CXL locally is comparable to accessing remote NUMA node memory. Additionally, as the proportion of write operations in the workload increases, the latency-bandwidth knee-point shifts left. Figures 4.4(g) and 4.4(h) show performance for read-only and write-only workloads under random access patterns. No significant performance differences were observed in these conditions.

4.2.4 Key Insights

Avoiding Remote Socket CXL Access.. CXL memory expansion is commonly used for memory-intensive applications, particularly those limited by memory capacity or bandwidth. In such cases, cross-socket memory access is not uncommon. However, developers should be aware of the performance drop when accessing CXL memory from a remote socket and avoid cross-socket CXL accesses where possible. Hardware vendors must also ensure compatibility between CXL memory modules and processors' CXL support through cooperative testing. With full CXL 1.1 support, we expect the bandwidth attainable when accessing CXL across sockets to approach that of MMEM across sockets.

Bandwidth Contention.. Previous research [206, 208] highlighted the impact of bandwidth contention. We further examined how memory latency changes with varying read-write ratios under bandwidth contention. Latency remains stable at low to moderate bandwidth utilization but increases sharply as utilization approaches higher levels, primarily due to queuing delays in the memory controller [207]. Additionally, when workloads include a higher proportion of write operations, the knee-point in latency occurs at lower memory bandwidth. While CXL memory is often described as a "tiered memory" solution, suitable only when MMEM is fully utilized [187, 201, 204], we argue against this view. Even if MMEM bandwidth is underutilized (e.g., by 30%), offloading part of the workload (e.g., 20%) to CXL memory can yield overall performance improvements. We recommend treating CXL memory as a valuable resource for load balancing, even when local DRAM bandwidth is not fully utilized. Further real-world evaluations support this insight (§4.4).

Comparison with FPGA-based CXL Implementations.. Intel recently disclosed performance metrics for their FPGA-based CXL prototype [187]. While they highlighted relative latency and bandwidth for soft and hard IP implementations, they did not share performance under load. Our measurements show that the ASIC CXL solution introduces only a $2.5\times$ latency overhead compared to MMEM, surpassing most of Intel's FPGA-based results. The FPGA-based solution achieved only 60% of PCIe bandwidth, while the Asteralabs A1000 prototype reached an impressive 73.6% bandwidth efficiency, clearly outperforming Intel's FPGA-based solution.



(a) Throughput of YCSB workload (b) Tail latency of YCSB-A (c) Tail latency of YCSB-C

Fig. 4.5: **KeyDB YCSB latency and throughput under different configurations.** (a) Average throughput of four YCSB workload under different system configuration. (b) Tail latency of YCSB-A (c) Tail latency CDF of YCSB-C, both reported by the YCSB client [211].

4.3 Memory Capacity-bound Applications

One of the most significant advantages of integrating CXL memory into modern computing systems is the potential for significantly larger memory capacities. To highlight the benefits, we focus on three specific use cases: (1) in-memory key-value stores, a commonly used application in data centers, (2) big data analytics applications, and (3) elastic computing from cloud providers.

4.3.1 In-memory Key-Value Stores

Redis [24] is a widely-used open-source in-memory key-value store and one of the most popular NoSQL databases. Redis employs a user-defined parameter, `maxmemory`, to limit memory allocation for storing user data. Similar to traditional memory allocators (e.g., malloc()), Redis may not release memory back to the system after key deletion, particularly when deleted keys reside on memory pages with active ones. As a result, memory provisioning must account for peak demand, making memory capacity a significant bottleneck for Redis deployments in data centers [212]. Google Cloud recommends keeping memory usage below 80% [213], while other sources suggest a 75% limit [212].

Due to the substantial infrastructure costs associated with memory-only deployment, Redis Enterprise [214], a commercial variant supported by leading cloud platforms (e.g., AWS, Google Cloud, Azure), introduces "Auto Tiering" [215], allowing data overflow to SSDs. This provides an economically viable solution for expanding database capacity beyond RAM limits. Given that Redis Enterprise is not available on our experimental platform, we use KeyDB as an alternative.

KeyDB extends Redis's capabilities by integrating KeyDB Flash, which uses RocksDB for persistent storage. The FLASH feature ensures all data is written to disk for persistence, while hot data remains in both memory and disk.

Methodology and Software Configurations

In this study, we explore the performance impact of maximizing memory utilization on a KeyDB server. We deploy a single KeyDB instance on a CXL-enabled server configured with seven *server-threads*. Unlike Redis's single-threaded model, KeyDB improves performance by allowing multiple threads to run the standard Redis event loop, effectively simulating several Redis instances in parallel. To minimize potential OS overhead, we disable Sub-NUMA Clustering (SNC) and Transparent Hugepages, and we enable memory overcommitting within the kernel. For KeyDB FLASH, all forms of compression in RocksDB are disabled to reduce software overhead.

Our empirical evaluation utilizes the YCSB benchmark, testing four distinct workloads:

1. **YCSB-A** (50% read, 50% update) for update-intensive scenarios,
2. **YCSB-B** (95% read, 5% update) for read-heavy operations,
3. **YCSB-C** (100% read) for read-only tasks,
4. **YCSB-D** (95% read, 5% insert) to simulate workloads accessing the most recent data.

These workloads are evaluated under different system configurations as detailed in Table 4.1. For consistency, we use "MMEM" to refer to main memory, distinguishing it from CXL memory. For configurations involving SSD spillover, we adjust the `maxmemory` parameter to match the portion of the workload expected to remain in memory. For Hot-Promote, we use `numactl` to distribute half of the dataset across CXL memory while limiting the main memory usage to half of the dataset size. The experiments utilize a key-value size of 1 KB, the YCSB default, with a Zipfian distribution for workloads A-C and the latest distribution for workload D. The total working set size is 512 GB.

Analysis

Figure 4.5 provides insights into the throughput variations across different configurations. Notably, regardless of the specific workload, running the entire workload on MMEM consistently delivers

Configuration	Description
`MMEM`	Entire working set in main memory.
`MMEM-SSD-0.2`	20% of the working set is spilled to SSD.
`MMEM-SSD-0.4`	40% of the working set is spilled to SSD.
`3:1`	Entire working set in memory (75% MMEM + 25% CXL, 3:1 interleaved).
`1:1`	Entire working set in memory (50% MMEM + 50% CXL, 1:1 interleaved).
`1:3`	Entire working set in memory (25% MMEM + 75% CXL, 1:3 interleaved).
`Hot-Promote`	Entire working set in memory (50% MMEM + 50% CXL), with hot page promotion kernel patches discussed in §4.1.

Table 4.1: **Configurations used in capacity experiments.**

the highest throughput. This result can be attributed to our workload being primarily constrained by memory capacity rather than memory bandwidth. The Hot-Promote configuration, which utilizes the Zipfian distribution to identify frequently accessed keys (hot pages) and migrate them from CXL to MMEM, performs nearly as well as running the workload entirely on MMEM. This highlights the effectiveness of the Hot-Promote approach in optimizing performance.

In contrast, interleaving data access between CXL and MMEM results in a noticeable performance decrease, with a slowdown of 1.2x to 1.5x compared to running the workload entirely on MMEM. This performance drop is primarily due to the higher access latency associated with CXL, as demonstrated in the tail latency plots for workload A and workload C (Figure **??**). The MMEM-SSD-0.2 and MMEM-SSD-0.4 configurations exhibit the poorest performance, showing a slowdown of approximately 1.8x compared to the pure MMEM solution and 1.55x compared to the CXL interleaving solution. This degraded performance is mainly due to the high access latency required to retrieve data from the SSD.

It is important to note that our choice of a Zipfian distribution ensures that the working set is largely cached in MMEM. If the keys were distributed uniformly, we would expect even worse performance due to the increased frequency of SSD accesses.

Insights

Our study demonstrates that the additional memory capacity provided by CXL can be a game-changer for applications like key-value stores, which are traditionally constrained by MMEM's limited capacity. Moreover, intelligent scheduling policies such as Hot-Promote further enhance the

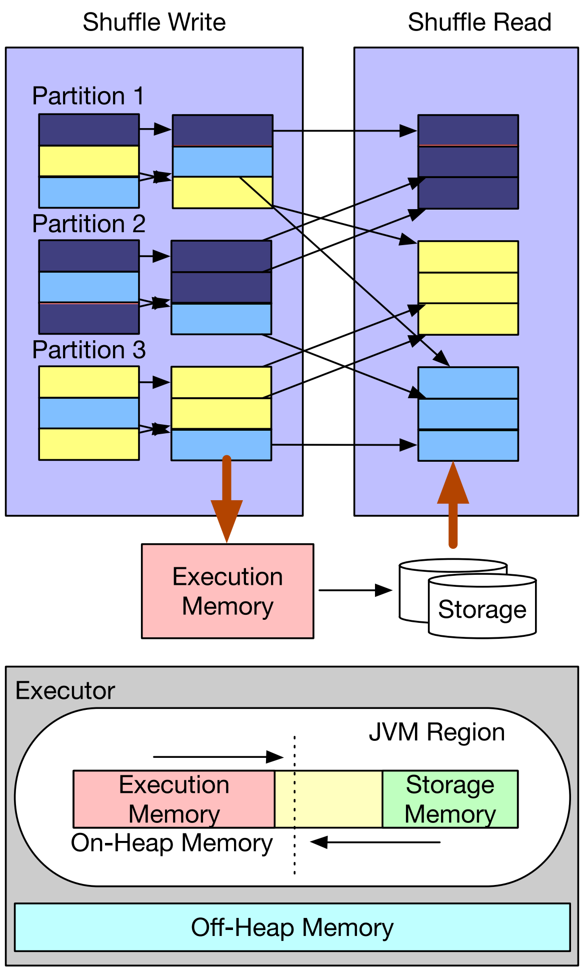
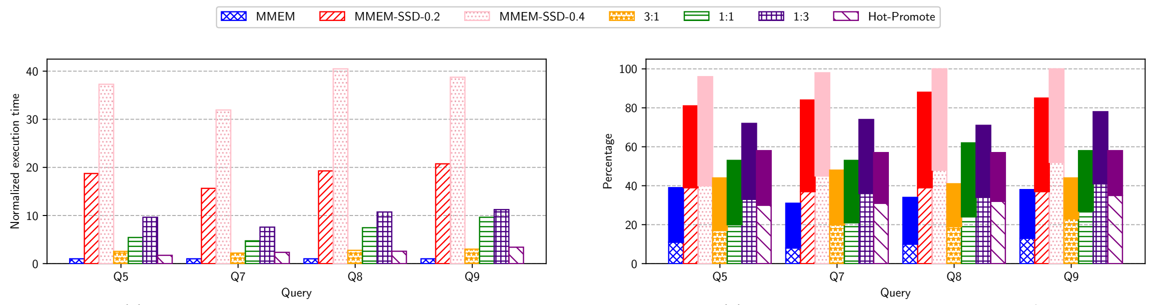


Fig. 4.6: **Spark memory layout and shuffle spill.** Each Spark executor possesses a fixed-size On-Heap memory, which is dynamically divided between execution and storage memory. If there is insufficient memory during shuffle operations, the Spark executor will spill the data to the disk.

benefits by optimizing performance across multiple memory types while also reducing operational costs.

4.3.2 Spark SQL

Big Data plays a crucial role in workloads managed by data centers. Due to the vast scale of data involved in Big Data analytical applications, memory capacity often becomes a bottleneck to performance [26]. Consider Spark [67], one of the most widely used Big Data platforms: A typical query requires shuffling data from multiple tables to process the next stage. Operations like *reduceByKey()* first partition data by key, then execute reduce operations on each key. This shuffling process involves significant disk I/O and network communication between nodes, introducing con-



(a) Normalized execution time of TPC-H queries (b) Percentage of time spent on shuffle r/w

Fig. 4.7: **Spark execution time and shuffle percentage.** (a) Execution time of each TPC-H query normalized to the execution time running on MMEM. (b) The percentage of time spent of shuffle operation for each query. The solid bars represent shuffle writes, while hollow bars represent shuffle reads.

siderable overhead to the query. In some cases, the performance of shuffling can dominate the overall workload performance [216].

During the shuffling process (Fig. 4.6), memory usage can exceed the available capacity or certain thresholds (e.g., `spark.shuffle.memoryFraction`). When this occurs, Spark can be configured to spill data to disk to avoid out-of-memory failures. However, since disk I/O is orders of magnitude slower than memory, this can significantly degrade the workload's performance.

Methodology and Software Configurations

In this experiment, we aim to evaluate whether we can reduce the number of servers required for a specific workload with minimal impact on overall performance. Thus, we compare the performance of Spark running TPC-H [217] on three servers without CXL memory expansion against two servers equipped with CXL memory expansion. We assume that the maximum amount of MMEM per server is 512 GB, so with three servers, we have a total of 1.5 TB of MMEM and 1 TB of CXL memory.

To trigger data spill within the workload, we configure Spark with 150 executors. Each Spark executor is allocated 1 core and 8 GB of memory, resulting in a total memory usage of 1.2 TB and 150 cores. We generate a 7 TB TPC-H dataset. The configuration settings detailed in Table 4.1 are applied as follows:

- **MMEM only**: We allocate 50 Spark executors and 400 GB of memory on each of the **three** servers. In this scenario, no data is spilled to disk, as each executor has sufficient memory.

- **MMEM/CXL interleaving**: We distribute the same number of executors (150) across the **two** CXL-enabled servers, which have 1 TB of CXL memory (512 GB from each of two CXL cards) and 1 TB of MMEM (512 GB each). For instance, in a configuration where MMEM and CXL memory usage is balanced (1:1 ratio), we allocate 75 executors to 600 GB of MMEM and 75 executors to 600 GB of CXL memory. In this case, data spill to disk is negligible.

- **Spill to SSD**: To simulate conditions where Spark executors run out of memory and need to spill data to SSD storage, we restrict memory allocation to either 80% or 60% of the total 1.2 TB MMEM, leading to approximately 320 GB and 500 GB of data being spilled to disk, respectively.

- **Hot-Promote**: Similar to the KeyDB experiment (§4.3.1), this configuration migrates hot data from CXL to MMEM.

We selected four TPC-H queries ($Q5$, $Q7$, $Q8$, and $Q9$), which are known for their intensive data shuffling demands [216], to evaluate our setup. Our measurements focus solely on the execution time of these queries, excluding data preparation and server setup times. SNC was disabled on all servers.

Analysis

Figure 4.7(a) illustrates variations in total execution time across different configurations. To provide a clear comparison, we normalized the total execution time against the best-case scenario, which is running the entire workload in MMEM. Similar to the KeyDB experiments, the interleaving approach results in a performance slowdown ranging from 1.4x to 9.8x compared to the optimal MMEM-only scenario, though it uses fewer servers. This performance degradation worsens as a larger proportion of memory is allocated to CXL. Nevertheless, it is crucial to note that even with this slowdown, the interleaving approach is significantly faster than spilling data to SSDs. Figure 4.7(b) shows that data shuffling exacerbates the total execution time due to intensified data spill issues.

A notable difference between the KeyDB and Spark experiments is the performance of the Hot-Promote configuration. While it performs well in KeyDB, the Spark SQL experiment reveals a more

Year	CPU	Max vCPU per server	Memory channels per socket	Max memory \TB	Required Memory (1 : 4) \TB
2021	IceLake-SP [218]	160	8xDDR4-3200	4	0.64
2022 (delayed)	Sapphire Rapids [219]	192	8xDDR5-4800	4	0.768
2023 (delayed)	Emerald Rapids [220]	256	8xDDR5-6400	4	1
2024+	Sierra Forest [221]	1152	12	4	4.5
2025+	Clearwater Forest [222]	1152	TBD	4	4.5

Table 4.2: **Intel Processor Series.**

than 34% slowdown compared to MMEM. Unlike the Zipfian distribution, which efficiently promotes hot keys from CXL to DDR, Spark SQL encounters considerable thrashing behavior within the kernel. Upon investigation, we traced the root cause to the hot page selection patch [203]. In its initial version, a sysctl parameter (`kernel.numa_balancing_promote_rate_limit_MBps`) was used to control the maximum promotion/demotion throughput. Later versions introduced an automatic threshold adjustment feature to balance the speed of promotion with migration costs. However, this automatic adjustment mechanism seems inadequate for the Spark SQL workload, which demonstrates reduced data locality and challenges the kernel's ability to efficiently promote frequently accessed pages. This issue is consistent with prior findings [187].

Insights

Our research indicates that utilizing CXL memory expansion offers a cost-effective solution for data center applications. A detailed theoretical examination of the Abstract Cost Model is postponed to §4.5. While the hot-promote patch shows significant advantages in key-value store workloads, its performance is notably lacking in Spark experiments. As system developers work to enhance software support for CXL within the kernel, they should proceed with caution. System-wide policies can have varied impacts depending on the specific characteristics of different applications.

4.3.3 Spare Cores for Virtual Machine

One widely-used application within Infrastructure-as-a-Service (IaaS) is Elastic Computing [223], where cloud service providers (CSPs) offer computational resources to users through virtual machines (VMs) or container instances. Given the diverse requirements of users, CSPs typically offer various instance types, each configured with different CPU cores, memory, disk, and network capacities. A commonly employed "optimal" CPU-to-memory ratio is $1 : 4$, as recommended by

AWS [224, 225]. For instance, an instance with 128 vCPUs would generally have 512 GB of DDR memory.

Advancements in server processor architecture and chiplet technology have rapidly increased the number of cores available in a single processor package, driven largely by CSPs' desire to lower per-core costs. Consequently, vCPU counts in 2-socket servers have increased from 160 to 256 over the past two years (Table 4.2), with projections reaching as high as 1152 vCPUs per server by 2025.

This surge in vCPUs exacerbates memory capacity bottlenecks, which are limited by DDR slot availability, DRAM density, and the cost of high-density DIMMs. For example, Intel's Sierra Forest Xeon supports up to 1152 vCPUs but is constrained by motherboard design to less than 4 TB of memory—falling short of the typical 4.5 TB required for VM provisioning [226]. This shortfall complicates the maintenance of a cost-effective vCPU-to-memory ratio, leading to underutilized vCPUs and revenue losses for CSPs. CXL memory expansion offers a solution by enabling memory capacity to scale beyond DDR limitations, thereby optimizing vCPU utilization and mitigating revenue losses for CSPs.

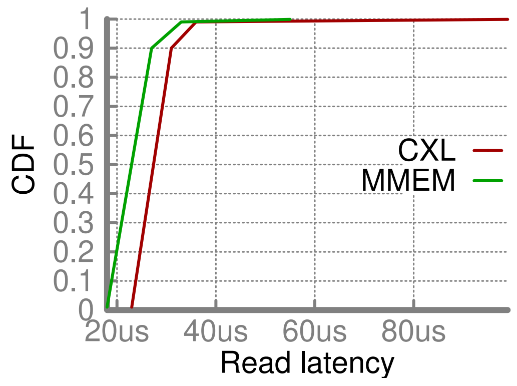
Methodology and Software Configurations

To evaluate the performance impact when an application runs entirely on CXL memory, we replicate the KeyDB configuration from earlier experiments (§4.3.1). Using *numactl*, we allocate the KeyDB instance exclusively to either MMEM or CXL memory. The workload for this evaluation is YCSB-C, which features 1 KB key-value pairs and a total dataset size of 100 GB. SNC is disabled in all configurations.

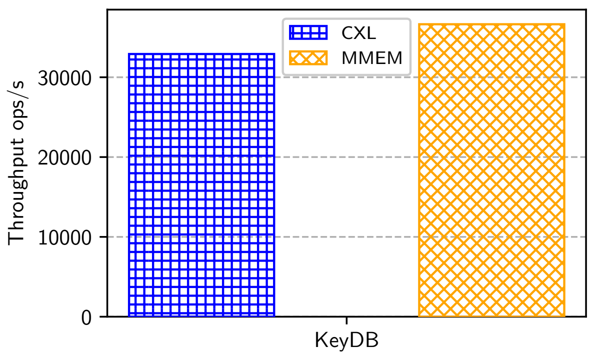
Analysis

The Cumulative Distribution Function (CDF) of read latency (Fig. 4.8(a)) shows that applications running on CXL memory experience a latency penalty of $9\% - 27\%$, which is less than the raw data fetching latency observed in §4.2. This difference is due to processing latency within Redis. Furthermore, the throughput of running the entire workload on CXL memory is approximately 12.5% lower than that of MMEM, as shown in Fig. 4.8(b).

Consider a server operating at a suboptimal vCPU-to-memory ratio of $1:3$:



(a) CDF of KeyDB YCSB-C



(b) Throughput of KeyDB YCSB-C

Fig. 4.8: **KeyDB Performance with YCSB-C on CXL/MMEM.**

1. Due to insufficient memory, only 75% of the available vCPUs can be sold at the optimal $1:4$ ratio, resulting in a 25% revenue loss. By implementing CXL memory expansion, CSPs can sell the remaining 25% of vCPUs at the optimal ratio.
2. Our benchmarks show that instances running on CXL memory perform 12.5% slower than those running on DDR memory for common workloads like Redis. Assuming a 20% price discount on such instances, CSPs can still recover approximately 80% of the lost revenue, resulting in a 27% improvement in total revenue ($20/75 = 26.77\%$).

Insights

Given the vast scale of Elastic Computing Service (ECS) applications in public clouds, the potential benefits of CXL memory expansion are considerable. However, maintaining an optimal virtual CPU (vCPU) to memory ratio, traditionally set at $1:4$, becomes increasingly complex with the rapid growth in processor cores. Although this ratio is a standard, its applicability in future cloud computing paradigms is being reevaluated. For example, Bytedance's Volcano Engine Cloud [227] demonstrates variability in resource allocation by offering different ratios: $1:4$ for general-purpose workloads, $1:2$ for compute-intensive tasks, and $1:8$ for memory and storage-intensive applications. The introduction of CXL memory expansion and pooling into these established ratios presents an intriguing area of exploration, raising important questions about the adaptability of cloud providers to evolving hardware capabilities and the subsequent effect on resource allocation standards.

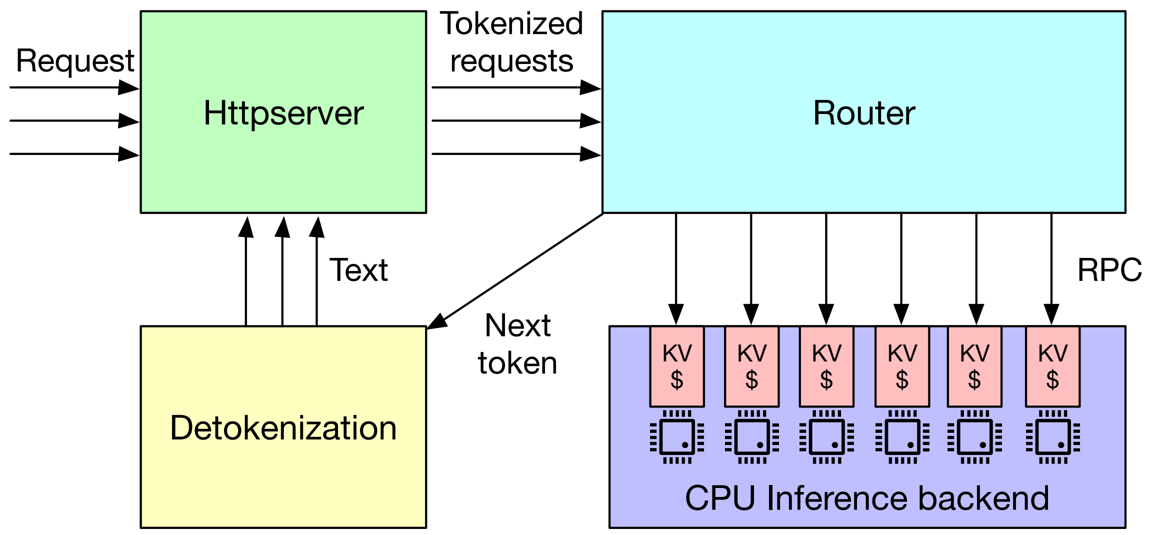


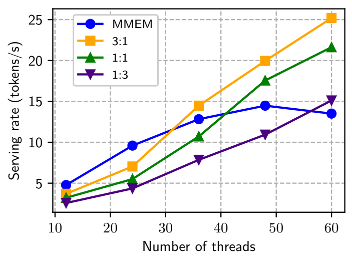
Fig. 4.9: **LLM inference framework.** The Httpserver receive requests and forward the tokenized requests to the CPU inference backend. The CPU inference backend serves the requests and reply the next token.

4.4 Memory Bandwidth-Bound Applications

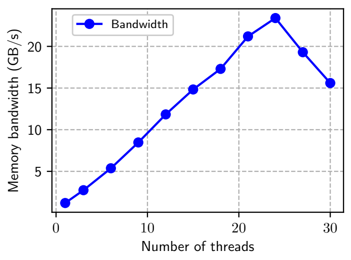
Another advantage of CXL memory expansion is its potential to provide additional memory bandwidth. We use Large Language Model (LLM) inference as an example to demonstrate how this can benefit real-world applications.

Recent research on LLMs [228] highlights that LLM inference is both memory-capacity and memory-bandwidth intensive. The limited capacity of GPU memory constrains the batch size of LLM inference jobs and reduces computational efficiency, as LLM models are highly memory-demanding. On the other hand, while CPU memory offers larger capacity, it suffers from lower bandwidth compared to GPU memory. The extra bandwidth and capacity offered by CXL memory make it a promising solution for alleviating these bottlenecks.

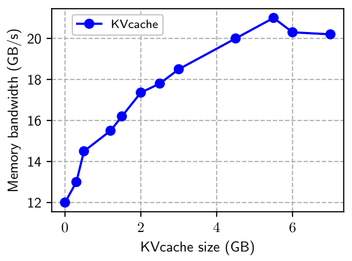
For instance, a CPU-based LLM inference job could benefit from the additional bandwidth provided by CXL memory. Similarly, a CXL-enabled GPU device could leverage the extra memory capacity from a disaggregated memory pool. Due to the current lack of CXL support in GPU devices, we focus on CPU-based LLM inference in our experiments to assess the potential impact of CXL memory's extra bandwidth. Moreover, since LLM inference applications are generally agnostic to the underlying memory technologies, our findings and implications should also apply to future CXL 2.0/3.0 devices.



(a) LLM inference serving rate vs. number of threads



(b) Memory bandwidth vs. number of threads for a single backend



(c) Memory bandwidth vs. KVcache size for a single backend

Fig. 4.10: **CPU LLM inference.**

LLM Inference Framework. Mainstream Large Language Model (LLM) inference frameworks, such as vLLM [229] and LightLLM [230], do not natively support CPU-based inference. Recently, Intel introduced the Q8chat LLM model [231], trained using their 4th Generation Intel Xeon® Scalable Processors. However, the inference code for Q8chat is not yet publicly available.

To address this gap, we developed an inference framework based on the open-source LightLLM framework [230], replacing its backend with a CPU inference backend. Figure 4.9 illustrates our implementation. In our framework, an HTTP server frontend receives LLM inference requests and forwards the tokenized requests to a router. The router distributes these requests to different CPU backend instances, each equipped with a Key-Value (KV) cache [232], a widely-used technique in LLM inference.

It is important to note that KV caching, despite its name, differs from traditional key-value stores in system architecture. KV caching occurs during multiple token generation steps within the decoder. During decoding, the model begins with a sequence of tokens, predicts the next token, appends it to the input, and repeats the process. This is how models like GPT [228] generate responses. The KV cache stores key and value projections used as intermediate data during decoding to avoid recomputation for each token generation. Prior research [232] has shown that KV caching is typically memory-bandwidth bound, as each sequence in the batch has its own unique KV cache, and different requests do not share this cache due to sequences being stored in separate contiguous memory spaces [233].

4.4.1 Methodology and Software Configurations

To investigate the benefits of CXL memory extension for applications with high memory bandwidth demands and limited MMEM bandwidth, we use an SNC-4 configuration to partition a single CPU into four sub-NUMA nodes. Each sub-NUMA node is equipped with two DDR5-4800 memory channels, which results in early memory bandwidth saturation at 67 GB/s (§4.2). We evaluate three different interleaving policies (3:1, 1:1, 1:3), as detailed in Table 4.1. The CPU inference backend is configured with 12 CPU threads, with memory allocation strictly bound to a single sub-NUMA domain. Each domain includes two DDR5-4800 channels and a 256 GB A1000 CXL memory expansion module via PCIe. By binding allocations to a single node, we ensure the early saturation of DDR5 channels.

For our experiments, we use the Alpaca 7B model [234], an extension of the LLaMA 7B model, which requires 4.1 GB of memory. The workload, derived from the LightLLM framework [230], includes a variety of chat-oriented questions. A single-threaded client machine on a baseline server sends HTTP requests with different LLM queries to simulate real-world conditions. The client ensures continuous operation of the CPU inference backends by maintaining a constant stream of requests. The prompt context is set to 2048 bytes to ensure a minimum inference response size. We progressively increase the number of CPU inference backends to monitor the LLM inference serving rate (measured in tokens/s).

4.4.2 Analysis

Figure 4.10(a) shows the inference serving rates across different memory configurations as the number of CPU inference backends increases. Initially, the serving rate improves almost linearly with available memory bandwidth. However, at 48 threads, MMEM bandwidth saturation limits the serving rate, while the interleaving configurations benefit from additional CXL bandwidth for continued scaling. With a high number of inference threads (60), a MMEM:CXL = 3:1 interleaving configuration outperforms the MMEM-only setup by 95%.

Among the interleaving policies, configurations with a higher proportion of data in main memory demonstrate better inference performance. Interestingly, we observe that beyond 64 threads, operating entirely on main memory is 14% less effective than a MMEM:CXL ratio of 1:3. This result is surprising given CXL's higher latency and lower memory bandwidth (§ 4.2). Figure 4.10(b) shows the memory bandwidth utilization, as measured by the Intel Performance Counter Monitor (PCM) [209], with increasing CPU thread counts. Bandwidth utilization grows linearly with thread count, plateauing at 24.2 GB/s for 24 threads. This trend allows us to estimate a bandwidth of approximately 63 GB/s at 60 threads, reaching 82% of the theoretical maximum. Our microbenchmark results (§4.2) suggest that this level of bandwidth utilization could cause significant latency spikes, corroborating the hypothesis that bandwidth contention plays a critical role in performance degradation.

Bandwidth contention may arise from loading the LLM model or accessing the KV cache. By adjusting the prompt context to infinity, the LLM model continuously generates new tokens for storage in the KV cache. Figure 4.10(c) illustrates the relationship between KV cache size and memory bandwidth consumption. The initial memory bandwidth of approximately 12 GB/s originates from I/O threads loading the model from memory. As the KV cache stores more tokens, memory usage increases linearly. However, bandwidth utilization plateaus around 21 GB/s.

4.4.3 Insights

Current tiered memory management in the kernel does not account for memory bandwidth contention. For a workload that utilizes a high percentage of MMEM bandwidth (e.g., 70%), existing page migration policies (§4.1) tend to move data from slower tiered memory (e.g., CXL) into MMEM, assuming sufficient memory capacity is available. As more data is written to MMEM, memory bandwidth utilization may rise to 90%, exponentially increasing access latency and causing a slowdown in the workload. This scenario is likely to occur in memory-bandwidth-bound applications, such as LLM inference. Therefore, the definition and management of tiered memory need to be reconsidered to address bandwidth contention effectively.

4.5 Cost Implications

Our comprehensive analysis in previous sections (§4.3, §4.4) demonstrates that the adoption of CXL memory expansion offers substantial benefits for data center applications, including comparable performance alongside significant operational cost savings. However, a major challenge in adopting

Parameter	Description
P_s	Throughput when (almost) entire working set is spilled to SSD on a server. Normalized to 1 in the cost model.
R_d	Relative throughput when the entire working set is in main memory on a server, normalized to P_s.
R_c	Relative throughput when the entire working set is in CXL memory on a server, normalized to P_s.
D	The MMEM capacity allocated to each server. For completeness only, not used in cost model.
C	The ratio of main memory to CXL capacity on a CXL server. E.g. 2 means the server has 2x MMEM capacity than CXL memory.
$N_{baseline}$	Number of servers in the baseline cluster.
N_{cxl}	Number of servers in the cluster with CXL memory to deliver the same performance as the baseline.
R_t	Relative TCO comparing a server equipped with CXL memory vs. baseline server. E.g. If a server with CXL memory costs 10% more than the baseline server, this parameter is 1.1.

Table 4.3: **Parameters of our Abstract Cost Model.**

innovative technologies like CXL lies in determining its Return on Investment (ROI). While detailed technical specifications and benchmark performance results are often available, accurately forecasting the Total Cost of Ownership (TCO) savings remains difficult. This is compounded by the complexity of running production-scale benchmarks and the limited availability of CXL hardware. Traditional cost models [235], which aim to provide such forecasts, typically require extensive internal data that is often sensitive and inaccessible. To address this challenge, we propose an Abstract Cost Model that estimates TCO savings without relying on internal or sensitive information. This model leverages a small set of metrics obtainable through microbenchmarks, along with empirical values that are easier to approximate or access, providing a viable approach to evaluating the economic feasibility of CXL technology adoption.

To illustrate the application of our Abstract Cost Model, we use a capacity-bound application (Spark SQL) as an example. However, this methodology can be extended to other types of workloads. For Spark SQL applications, the additional memory capacity provided by CXL reduces the amount of data spilled to SSD, resulting in improved throughput and performance. This, in turn, means that fewer servers are required to meet the same performance target.

Given that the workload maintains a relatively consistent memory footprint (i.e., the size of the active dataset) during execution, we can approximate the execution time of the workload by dividing it into three distinct segments:

1. The segment processed with data stored in MMEM,

2. The segment processed with data stored in CXL memory,

3. The segment processed with data offloaded to SSD storage.

We begin by collecting the following measurements from microbenchmarks on a single server:

- **Baseline performance** (P_s): Measure the throughput when (almost) the entire working set is spilled to SSD. While the absolute number is not directly used in our cost model, we normalize this value to 1 for comparison.
- **Relative performance with the entire working set in MMEM** (R_d): Using the same workload, measure the throughput when the entire working set resides in MMEM. Normalize this value to P_s to express the relative performance improvement (i.e., how much faster compared to the baseline).
- **Relative performance with the entire working set in CXL memory** (R_c): Using the same workload, measure the throughput when the entire working set resides in CXL memory. Normalize this value to P_s to express the relative performance compared to the baseline.

We then formulate our cost model using the parameters summarized in Table 5.1. For a working set size of W, the execution time of the baseline cluster can be approximated as the sum of two segments: (1) the segment executed with data in MMEM and (2) the segment executed with data spilled to SSD.

$$T_{baseline} = \frac{N_{baseline}D}{R_d} + (W - N_{baseline}D)$$

The execution time of the cluster with CXL memory could be approximated in a similar way. It includes the segment that is executed with data in main memory, in CXL memory, and spilled to SSD respectively.

$$T_{cxl} = \frac{N_{cxl}D}{R_d} + \frac{N_{cxl}D}{CR_c} + (W - N_{cxl}D - \frac{N_{cxl}D}{C})$$

To meet the same performance target, $T_{baseline} = T_{cxl}$:

$$\frac{N_{baseline}D}{R_d} - N_{baseline}D = \frac{N_{cxl}D}{R_d} + \frac{N_{cxl}D}{CR_c} - N_{cxl}D - \frac{N_{cxl}D}{C}$$

With some simple transformations, we get the ratio between N_{cxl} and $N_{baseline}$:

$$\frac{N_{cxl}}{N_{baseline}} = \frac{CR_c(R_d - 1)}{R_cR_d(C+1) - CR_c - R_d}$$

TCO saving can then be formulated as follows.

$$TCO_{saving} = 1 - \frac{TCO_{cxl}}{TCO_{baseline}} = 1 - \frac{N_{cxl}R_t}{N_{baseline}}$$

For example, suppose $R_d = 10$, $R_c = 8$, $C = 2$, we get $\frac{N_{cxl}}{N_{baseline}} = 67.29\%$ from the cost model. This means that by using CXL memory, we may reduce the number of servers by 32.71%. And if we further assume $R_t = 1.1$ (a server with CXL memory costs 10% more than the baseline server), the TCO saving is estimated to be 25.98%.

Our Abstract Cost Model provides an easy and accessible way to estimate the benefit from using CXL memory, providing important guidance to the design of the next-generation infrastructure.

Extending Cost Model for more realistic scenarios. In line with previous research [235], our Abstract Cost Model is designed to be adaptable, allowing for the inclusion of additional practical infrastructure expenses such as the cost of CXL memory controllers, CXL switches (applicable in CXL 2.0/3.0 versions), PCBs, cables, etc., as fixed constants. However, a notable constraint of our current model is its focus on only one type of application at a time. This becomes a challenge when a data center provider seeks to evaluate cost savings for multiple distinct applications, each with unique characteristics, especially in environments where resources are shared (for instance, through CXL memory pools). This scenario introduces complexity and presents an intriguing challenge, which we acknowledge as an area for future investigation.

4.6 Related Work

As memory-intensive applications like machine learning and High-Performance Computing (HPC) continue to grow, expanding memory capacity and bandwidth has become a critical challenge [236–238]. Compute Express Link (CXL) [10, 185, 206] has emerged as a promising solution, offering a novel interconnect technology that connects external memory devices via PCIe, significantly enhancing both memory capacity and bandwidth.

Research on CXL technology is extensive. Studies such as [185, 187, 206] explore CXL's potential for resource disaggregation and memory expansion, highlighting its ability to address memory bottlenecks in large-scale applications. However, much of the research relies on simulations [37, 206] or FPGA-based CXL hardware [187, 197, 199], limiting practical insights into production-ready ASIC-based hardware. More recent empirical evaluations, such as [187, 196], have begun to investigate the performance of ASIC-based CXL hardware, offering valuable data on its real-world potential. These studies show that while CXL introduces higher latency compared to local memory, this gap narrows for cross-socket memory access, making CXL a feasible option for memory tiering.

Other work, such as [37, 189, 197], has examined the trade-offs between performance and cost in CXL systems, particularly through synthetic benchmarks. For instance, [189] investigates kernel-level techniques for promoting hot pages from slower to faster memory tiers to boost performance while maintaining application transparency. Similarly, [235] introduces cost models for memory pooling in CXL, but further research is needed to evaluate the economic viability of migrating specific workloads to CXL-enabled systems.

4.7 Summary

In this chapter, we provide a comprehensive empirical evaluation of Compute Express Link (CXL) in real-world data center applications, filling a critical knowledge gap left by prior theoretical studies. Our findings reveal both the potential and limitations of CXL, offering actionable recommendations for its ongoing development to better serve data-centric computing environments. Based on our benchmarks, we also develop an Abstract Cost Model that can estimate the TCO savings without relying on internal or sensitive data, providing important guidance to the design of our next generation infrastructure

Chapter 5

Conclusions and Future Work

In this dissertation, we take a top-down approach and explore the optimal memory management solutions for three important layers in the cloud stack, i.e. Service, OS and Hardware.

5.1 Future Work

We next discuss problems that we leave open in this dissertation and on-going work.

5.1.1 CXL-based KV Cache Storage

Autoregressive large language models (LLMs) generate output tokens sequentially, where the generation of each token involves the attention computation using key-value (KV) of its preceding tokens [239–241]. This sequential dependency makes LLM inference both compute- and memory-intensive. LLM inference typically includes two stages: the prefill stage, where all input tokens are processed to generate the initial output token, and the decode stage, where the rest of the output tokens are generated one by one until the model generates an end-of-sequence token [242–244].

For applications such as chatbot and coding assistant, LLM serving systems aim to minimize the time to finish the prefill stage, or time to first token (TTFT). In production, service-level objective (SLO) for TTFT is typically 400ms [243]. To meet such SLO, LLM serving systems often cache the previously-computed KV data of the preceding tokens (i.e., prefix) in GPU memory, to avoid re-computing them for future requests that have the same prefix [243, 245, 246]. Storing KV cache reduces the overall computational load and significantly improves throughput by trading memory for computation.

In production chatbot applications that support large context windows, the demand for KV cache storage grows rapidly by the number of inference requests from users, which cannot be fully accommodated by the limited and expensive GPU memory [247]. Researchers thus developed techniques to offload KV cache to CPU memory, leveraging the larger CPU memory capacity to reduce GPU memory pressure [246, 248, 249]. However, as larger LLMs and support for long-context inference requests continue to emerge, the approach of storing KV cache to CPU memory is still insufficient. For example, in LLaMA-2-7B, KV cache of token in FP32 precision is 1024KB; KV cache of a single request with 4096 tokens (maximum context length) is 4GB [250]. The memory demand from serving many concurrent long-context requests can easily overwhelm even high-end memory servers [245, 251].

Practitioners increasingly turn to more scalable memory architectures, such as Compute Express Link (CXL) memory [37, 188, 252], to address the growing memory demands of large-scale systems. CXL expands memory capacity by connecting additional DRAM to servers via PCIe, while maintaining low-latency access. It offers a promising solution to the KV cache storage demand in LLM serving.

In this paper, we propose to leverage CXL memory for storing KV cache, with the goal to improve serving throughput while retaining SLO on TTFT, and reduce KV cache storage pressure for the upper-level LLM serving system. This paper makes the following contributions:

- We present the first measurement of CXL-GPU interconnect and evaluate its feasibility for KV cache storage. We show that the data-transfer latency and bandwidth on CXL-GPU interconnect is on par with CPU-GPU interconnect.

- We present our design of CXL-based KV cache storage interface and evaluate its performance improvement to LLM serving, on our platform that is the first to successfully integrate ASIC-CXL device and GPU. Our results show competitive TTFT achieved by CXL-based prefix caching.

- We examine the cost-efficiency in using CXL for KV cache storage in production via Return on Investment (ROI) modeling. Estimates show a promising reduction in GPU compute cost when using CXL for KV cache storage. We also identify promising future research directions.

We now present the design and implementation of our CXL-based KV cache storage interface for LLM serving. We also describe the hardware platform used to evaluate our design.

Design and implementation. Our goal is to develop a CXL storage interface, named *KVExpress*, which can be integrated into existing LLM serving systems for saving and loading KV cache of inference requests. *KVExpress* provides two external APIs to its upper-level serving system: `save` and `load`. The `save` takes a unique identifier of a token chunk as input, and copies its KV cache from GPU to CXL memory. The `load` takes a unique identifier of a token chunk as input, and finds if its KV cache exists in CXL memory, if so, copies the KV cache from CXL memory to GPU. A token chunk can consist of one or more tokens. The unique identifier of a token chunk t_i for a sequence is the hash of the content of t_i and the hash of its prefix $\langle t_0, ..., t_{i-1} \rangle$. If the prefix of a sequence of a current request has been computed and saved into CXL, *KVExpress* will load the KV cache of the prefix from CXL and use it when computing for this request [245].

To avoid calling `save` and `load` too frequently and incurring unnecessary overhead to the upper-level serving system, `save` is called only when a request is finished so the KV cache of all the tokens for that request is saved at once; `load` is called for a request prior to its prefill computation.

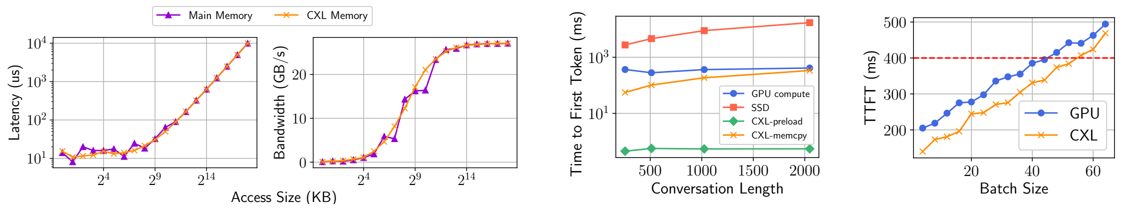
We implement our design of *KVExpress* in gpt-fast [253], a simple and low-latency text generation system with support on a number of widely-used inference optimizations [254–256] and open-source LLMs [250, 257]. We further modify gpt-fast to support our evaluation on batched inference.

Hardware platform. Our single socket server is equipped with Intel Xeon Platinum processors [258], 1TB of 4800 MHz DDR5 memory, an NVIDIA H100 GPU with 96GB HBM, and a CXL memory expansion card with 256 GB of DDR5 memory at 4800 MHz [188]. While prior works [259–261] have explored utilizing CXL for accelerators, to our knowledge, our work is the first implementation to successfully integrate a real ASIC-CXL device and a GPU within a single inference server.

5.2 Performance Evaluation

In Section 5.2.1, we measure the latency and bandwidth of CXL-GPU interconnect for data transfer to assess the feasibility of storing KV cache on CXL devices. In Section 5.2.2, we compare the

TTFT of KV re-compute, prefix caching with CXL, and prefix caching with GPU, to understand if *KVExpress* can achieve similar TTFT as existing approaches for prefill requests under varying context lengths. In Section 5.2.3, we study the maximum batch size achieved while retaining a given SLO on TTFT between KV re-compute and prefix caching with CXL.



(a) CPU-GPU/CXL-GPU interconnect. (b) TTFT comparison. (c) Max BS under SLO.

Fig. 5.1: **Experiment results.** (a) Latency and bandwidth measurements across different access sizes, CXL-GPU interconnect performs similarly as CPU-GPU interconnect. (b) TTFT comparison between KV re-compute and prefix caching with CXL or GPU. (c) Serving throughput comparison under a fixed SLO constraint (400ms).

5.2.1 Measurements on CXL-GPU interconnect performance

KV cache storage requires low-latency access (e.g., from host memory to GPU memory). Although prior studies [188,252] show that accessing CXL memory from the host CPU is over $2\times$ slower than accessing local memory, none of their measurements involves any interaction with the GPU. In this paper, we evaluate the performance characteristics of the CXL-GPU interconnect by measuring the latency and bandwidth of copying data from CXL memory to the GPU. Transferring in the reverse direction yields similar results [262]. Since CXL memory devices are exposed to the system as NUMA nodes without CPUs by default [188], we allocate a set of host buffers on the CXL NUMA node and use `cudaMemcpyAsync` to copy data between the host buffers and GPU device buffers allocated via the CUDA API [263]. We evaluated transferring data of sizes ranging from 1KB to 256MB.

Figure 5.1(a) shows our experiment results: the performance of the CXL-GPU interconnect **is unexpectedly on par with** traditional CPU-GPU memory transfers, exhibiting no significant slowdown. Latency remains low for smaller access sizes but increases exponentially once the size exceeds 64KB. Meanwhile, bandwidth increases almost linearly with data size and saturates around 4MB. This indicates that, while the CPU oversees the data transfer, the data path actually bypasses the host's local memory, flowing directly from CXL memory to GPU buffers via PCIe. Our results demonstrate that the CXL-GPU interconnect operates efficiently with minimal latency overhead, positioning it as a promising expansion for KV cache storage in addition to CPU memory.

5.2.2 Evaluation on TTFT under varying input context length

Given that CXL-GPU interconnect performs nearly the same as CPU-GPU interconnect, we further study if CXL-based KV cache storage can achieve similar TTFT as existing approaches in completing the prefill stage computation for an inference request. We evaluate three approaches:

- **KV re-compute:** Compute KV data of all input tokens for the request with GPU.
- **Prefix caching with CXL:** Load KV cache of the prefix tokens for the request from CXL to GPU.
- **Prefix caching with GPU:** Store and use KV cache in GPU for the prefix tokens for the request.

We measure the TTFT of the aforementioned approaches on conversation requests of input length ranging from 256 to 2048 tokens from the ShareGPT-Vicuna-Unfiltered dataset [264]. We use the LLaMA-2-13B as the underlying model for our evaluation. Figure 5.1(b) shows the TTFT (y-axis in log-scale) achieved by the evaluated approaches for requests of varying input context length (x-axis).

Compared to the other approaches, prefix caching with GPU (denoted as "PC-GPU" in Figure 5.1(b)) achieves the smallest TTFT (0.44ms to 0.56ms) constantly across different input context lengths. Such performance is expected as there is no data transfer latency and computation of KV data is only needed for tokens after the prefix. This approach is an optimal baseline that is however difficult to achieve in practice due to limited memory capacity of existing GPU models and the rapidly growing demand of KV cache storage in LLM serving.

Comparing prefix caching with CXL (denoted as "PC-CXL") and KV re-compute, prefix caching with CXL performs at least as good as computing KV data on GPU from scratch. Prefix caching with CXL achieve TTFT ranging from 55ms to 336ms, with slight increase in latency as input size length grows. The close performance gap between storing prefix KV cache in CXL memory and full KV re-computation indicates that there is a potential opportunity to reduce GPU compute cost with adaptation of CXL devices for memory capacity expansion in LLM inference.

5.2.3 Evaluation on serving throughput while adhering SLO

By storing the KV cache of the inference request prefix in CXL memory and thus reducing re-computation during the prefill stage, we can effectively reduce the computational load on the GPU. The saved GPU compute can be re-allocated to handle a larger number of concurrent inference requests. In other words, the LLM serving system can achieve a higher serving throughput, by handling a larger batch size of inference requests using the saved GPU compute, while maintaining the same SLO on TTFT [243].

Figure 5.1(c) shows the TTFT achieved by KV re-compute and prefix caching with CXL under varying batch size. The horizontal red-dashed line indicates our SLO limit–the maximum TTFT that can be tolerant in production. The typical SLO is 400ms used for LLaMA-2 [265]. As shown in Figure 5.1(c), with KV re-compute, the evaluated serving system (§**??**) can handle a maximum batch size of 44 before hitting the SLO limit. On the other hand, when leveraging CXL for storing KV cache, the system can handle a maximum batch size of 57, which is a **30% increase** compared to KV re-compute. Our initial evaluation on SLO-adhering serving throughput highlights the performance benefits of utilizing CXL memory for KV cache storage, particularly in scenarios that require efficient scaling under strict latency requirements.

5.3 Cost-Efficiency Modeling

We develop a model to estimate the Return on Investment (ROI) of deploying *KVExpress* in production. Conceptually, each prefill request consists of two distinct parts: 1) Loading KV cache data for the prefix (i.e., the history context) from CXL memory; 2) Performing computation on the new prompt (i.e., the follow-up prompt in multi-round conversations).

By replacing computation with memory accesses, we reduce the overall computational load, thereby lowering the demand for FLOP/s while still meeting the same SLO. This results in significant cost savings for LLM inference (Figure 5.2):

- **Assumption:** Assume a GPU has a computational power of 100 TFLOP/s, an average prefill request requires 25 TFLOP of computation, and the SLO for prefill is 0.5 seconds.

- **Baseline:** To complete the prefill request within the SLO, each request demands 50 TFLOP/s

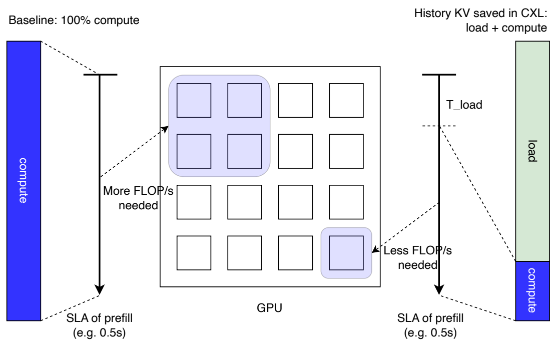

Baseline: 100% compute
compute
More FLOP/s needed
SLA of prefill (e.g. 0.5s)
GPU
Less FLOP/s needed
SLA of prefill (e.g. 0.5s)
T_load
History KV saved in CXL: load + compute
load
compute

Fig. 5.2: Example of ROI modeling: replace computation with memory access

(25 TFLOP/0.5s), meaning a single GPU can serve 2 prefill requests.

- **KVExpress:** By spending 0.1s loading KV cache data of the history context, we reduce the computational demand to 2.5 TFLOP (assuming the new prompt accounts for 10%). To meet the same SLO, the remaining computation must be finished within 0.4s, requiring 6.25 TFLOP/s (2.5 TFLOP/0.4s). In this case, a single GPU can serve 16 prefill requests, yielding an **8x improvement over the baseline**.

This allows for a reduction of 87.5% in the number of GPUs required for the same prefill SLO, resulting in substantial cost savings for LLM inference applications (more details in Appendix [**?**]).

Table 5.1: ROI Modeling

C_0	Avg. FLOPs needed by a prefill request in an initial request. Can be estimated as $C_0 = 2ML$, where M is the model parameters and L is the avg. sequence length.
C_1	FLOPs needed by new prompt in a follow-up request. Can be estimated as $C_1 = rC_0$, where r is the avg. ratio of the new prompt (e.g., 10%).
T_{slo}	SLO of prefill (e.g., 0.5s).
T_{load}	Avg. time to load KV cache from memory (e.g., 0.1s).
P	Computation power (FLOP/s) of the GPU.
P_0	FLOP/s needed for the initial request. $P_0 = C_0/T_{slo}$
P_1	FLOP/s needed for the new prompt. $P_1 = C_1/(T_{slo} - T_{load})$
R_{gpu}	Request per second (RPS) a single GPU can support. $R_{gpu} = P/(P_0(1-h) + P_1 h)$, where h is the ratio of multi-round requests.
N_{cxl}	Number of GPUs needed using our CXL memory scheme. $N_{cxl} = \lceil R/R_{gpu} \rceil = \lceil \frac{R}{P/(P_0(1-h)+P_1 h)} \rceil$
$N_{baseline}$	Number of GPUs needed without any KV cache stored (i.e., all data discarded after prefill). $N_{baseline} = \lceil \frac{R}{P/P_0} \rceil$

5.4 The Potential of Memory Disaggregation for AI/ML Workloads

Storing KV cache in GPU memory for LLM inference can quickly lead to memory saturation, limiting serving scalability and performance. KV cache storage on CPU memory becomes limited as model size and request context length increase. To that extent, we explore CXL memory for KV cache offloading, in which CXL offers expanded capacity with low-latency access. Our preliminary results show that CXL-CPU data transfer has similar latency and bandwidth as the CPU-GPU counterpart. In addition, CXL-based KV cache offloading provides similar performance compared to full KV re-compute on GPUs, while supporting larger workloads. Specifically, using CXL memory for KV cache storage increased the maximum batch size by 30%, while maintaining the same SLO on TTFT. Our cost-efficiency analysis further shows the potential for using CXL memory to substantially reduce the GPU compute cost for high-throughput LLM serving under SLO. Looking ahead, future work will explore the integration of CXL memory with multi-GPU systems, focusing on maintaining cache coherence across GPUs that could further enhance the scalability and efficiency of LLM inference.

Appendix A

Appendix

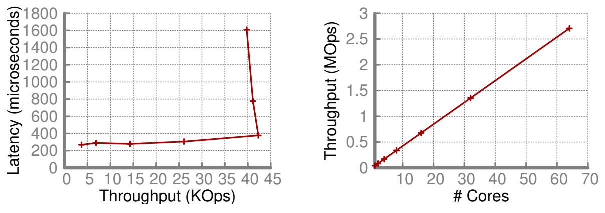
A.1 Jiffy: Additional Evaluation

We now present additional results for Jiffy, including: an evaluation of its control plane (Appendix A.1.1), and sensitivity analysis for various system parameters (Appendix C.3).

A.1.1 Controller Performance

Jiffy adds several components at the controller compared to Pocket, including all of metadata management, lease management and handling requests for data repartitioning. As such, we expect its performance to be lower than Pocket's metadata server. We deem this to be acceptable as long as it can still handle control plane request rates typically seen for real world workload, *e.g.*, a peak of a few hundred requests per second, including lease renewal requests, for all of our evaluated workloads and those evaluated in [42].

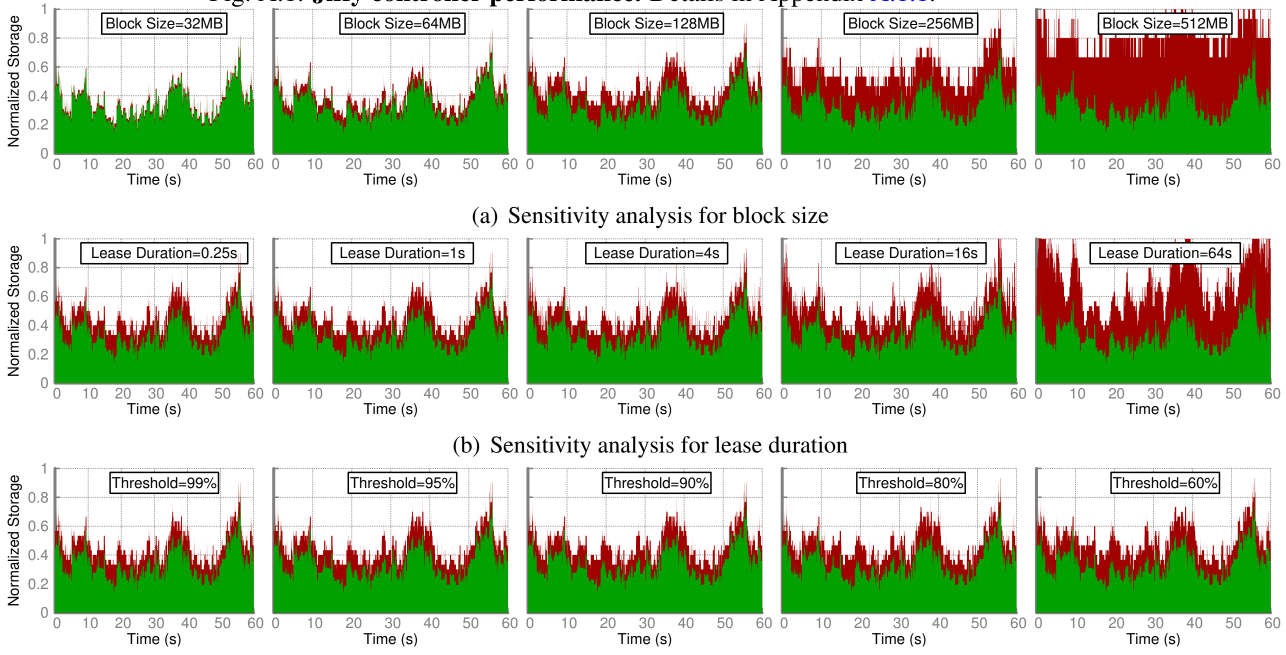
Figure A.1(a) shows the throughput-vs-latency curve for Jiffy controller operations on a single CPU core of an m4.16xlarge EC2 instance. The controller throughput saturates at roughly 42 KOps, with a latency of 370us. While this throughput is lower than Pocket ($\sim$ 90KOps per core), it is more than sufficient to handle control plane load for real-world workloads. In addition, the throughput scales almost linearly with the number of cores, since each core can handle requests independent of other cores for a distinct subset of virtual address hierarchies (Figure A.1(b)). Moreover, the Jiffy control plane readily scales to multiple servers by partitioning the set of virtual address hierarchies across them.



(a) Controller throughput vs. latency on a single CPU core.

(b) Controller throughput scaling with multiple cores.

Fig. A.1: **Jiffy controller performance.** Details in Appendix A.1.1.



(a) Sensitivity analysis for block size

(b) Sensitivity analysis for lease duration

(c) Sensitivity analysis for (high) repartition threshold

Fig. A.2: **Jiffy sensitivity analysis** for (a) block size (b) lease duration and (c) repartition threshold for the file data structure. Green area corresponds to used capacity, while red area corresponds to allocated capacity under Jiffy. See Appendix C.3 for details.

A.1.2 Sensitivity Analysis

We now perform sensitivity analysis for various system parameters in Jiffy, including block size (§2.1.1), lease duration (§2.1.2) and thresholds for data repartitioning (§2.1.3). We use files as our underlying data structure, and use the Snowflake workload from Figure 2.1. These results can be contrasted directly with Figure 2.10(a) (center), which corresponds to our default system parameters (128MB blocks, 1s lease duration and 95% of block occupancy as repartition threshold). For each parameter that we vary, the other to remain fixed at their default values.

Block size (Figure A.2(a)). As discussed in §2.1.1, the block size in Jiffy exposes a tradeoff between the amount of metadata that needs to be stored at the control plane, and resource utilization.

This is confirmed in Figure A.2(a), where increasing the block size from 32MB to 512MB increases the disparity between allocated and used capacity, and therefore decreases the resource utilization. The default block size in Jiffy is set to 128MB for two main reasons: (1) it allows high enough utilization with low enough metadata overhead (a few megabytes for even thousands of gigabytes of application data), and (2) it is the default block size used in existing data analytics platforms; as such, 128MB blocks ensure seamless compatibility with such frameworks.

Lease duration (Figure A.2(b)). As shown in Figure A.2(b), lease duration in Jiffy controls resource utilization over time. As we increase lease durations from 0.25 seconds to 64 seconds, resource utilization increases since Jiffy does not reclaim (potentially unused) resource resources from jobs until their leases expire. At the same time, if we keep lease duration too low, applications would renew leases too often, resulting in higher traffic to the Jiffy controller. We find a lease duration of 1s to be a sweet spot, ensuring high enough resource utilization, while ensuring the number of lease requests for even thousands of concurrent applications is only a few thousand requests per second — well within Jiffy controller's limits on a single CPU core.

Repartition threshold (Figure A.2(c)). Finally, Figure A.2(c) shows the impact of (high) repartition threshold on resource utilization. As expected, lowering the repartition threshold leads to poor utilization, since it triggers pre-mature allocation of new blocks to most files in our evaluated workload. However, since the size of the block (128MB) is much smaller than the amount of data written to each file in the workload (often several gigabytes), this overhead is relatively small when compared to effect of other parameters. However, a larger value of high repartitioning threshold results in more frequent block allocation requests to the controller; we find that our default value of 95% provides a reasonable compromise between resource utilization and number of control plane requests.

Supplementary Materials

A. Multiplexing $M + N$ Iterator Executions for Maximizing Pipeline Utilization

We claimed in §3.3.3 that if $t_c = \eta \cdot t_d$ for all offloaded iterator executions, it is always possible to multiplex $m + n$ concurrent iterator executions and fully utilize all memory and logic pipelines. We prove our claim by providing a staggered scheduling algorithm (Algorithm 1) that ensures such multiplexing across $m + n$ iterator executions. The scheduler processes $m + n$ iterator execution requests, assigning each a memory pipeline, a logic pipeline, and staggered start times. These requests are then executed in the respective memory pipelines. Through this staggered scheduling approach, Jiffy fully utilizes the n memory pipelines and m logic pipelines, ensuring no resources are wasted. Note that this algorithm is a simplified version to illustrate the potential for full pipeline saturation under the given condition. Jiffy's scheduler implements a real-time algorithm to multiplex incoming requests on the fly.

Algorithm 1 Staggered-Scheduling

```
1: $m, n \leftarrow$ number of logic, memory pipelines
2: $L_i, M_j \leftarrow$ $i^{th}$ logic pipeline, $j^{th}$ memory pipeline
3: $t_d \leftarrow$ data fetch time per pointer traversal iteration
4: while true do
5:     Dequeue $n + m$ requests from network stack
6:     for $i \leftarrow 1$ to $m + n$ do
7:         Assign request $R_i$ to ($M_{i \bmod n}$, $L_{i \bmod m}$)
8:         Schedule $R_i$ to start at time $(i - 1) \cdot \frac{t_d}{n}$
9:     Start requests as scheduled at memory pipelines
```

A.1 PULSE Empirical Analysis

Prior studies have shown that real-world data-centric cloud applications spend a significant fraction of time traversing pointers, as summarized in Fig. A.1.

B. PULSE Supported Data Structures

We adapt 13 data structures across 4 popular open-sourced libraries to PULSE's iterator abstraction (§??). In particular, we outline how the data structure implementations for certain operations can be expressed using `init()`, `next()`, and `end()`. For simplicity and readability, (i) we assume that the data structure developer defines a macro, `SP_PTR_(variable_name)`, as the address of the variable resides on the `scratch_pad`, and (ii) we omit obvious type conversions for de-referenced pointers.

We analyze two widely used categories of data structures: lists and trees. In our analysis, we find that the top-level data structure APIs (i.e., the APIs used by applications) use the same base function under the hood. For instance, list and forward list in the STL library share the same internal function, `std::find()`. We summarize our findings in Table A.1, including the data structure libraries, their category, the top-level data structure APIs, and the internal base function.

List structures. Our surveyed list structures already follow the execution flow of PULSE iterator: `init()`, `next()`, and `end()`.

These data structures generally have compute-intensive `end()` functions to check multiple termination conditions, while their `next()` function simply dereferences a single pointer to the next node. Listing A.1 and Listing A.2 demonstrate a linked list with two termination conditions: (i) value is found or (ii) search reaches the end. To indicate which condition is met, a special flag (*e.g.*, `KEY_NOT_FOUND`) is written on the `scratch_pad`. Listing A.3 and Listing A.4 describe a bitmap that uses a hashtable internally, where colliding entries are stored in linked lists within the same bucket. As such, the PULSE iterator interface resembles that of `std::list` quite closely.

Tree-like data structures. Compared to list structures, tree data structures require more computation in the `next()` function, as the next pointer is determined based on the value in the child node.

Application	**% of time spent in pointer traversal**
GraphChi [94]	$\sim 93\%$
MonetDB [102]	$70\% - 97\%$
GC in Spark [67]	$\sim 72\%$
VoltDB [103]	Up to 49.55%
MemC3 [104]	Up to 21.15%
DBx1000 [105]	$\sim 9\%$
Memcached [106]	$\sim 7\%$

(a) Survey from prior studies

Fig. A.1: **Time cloud applications spend in pointer traversals** based on prior studies

For instance, in `Btree` (Listing A.5, A.6), the next function iterates through internal node keys, comparing them to the search key. Interestingly, `std::map` (Listing A.7, A.8) and Boost AVL trees (Listing A.9, A.10) share the same offload function structure, with only minor implementation and naming differences.

Table A.1: Additional data structure supported by PULSE.

Data Structure	**Category**	**Library**	**Data structure API**	**Internal function**	**Original code**	**PULSE code**
List	List	STL	`std::find(start, end, value)`	`std::find(start, end, value)`	Listing A.1	Listing A.2
Forward list						
Bimap		Boost	`find(key, hash)`	`find(key, hash)`	Listing A.3	Listing A.4
Unordered map						
Unordered set						
Btree	Tree	Google	`find(&key)`	`internal_locate_plain_compare(key, iter)`	Listing A.5	Listing A.6
Map		STL		`_M_lower_bound(x, y, key)`	Listing A.7	Listing A.8
Set						
Multimap						
Multiset						
AVL tree		Boost		`lower_bound_loop(x, y, key)`	Listing A.9	Listing A.10
Splay tree						
Scapegoat tree						

B.1 List data structure in STL library

Listing A.1: C++ STL realization for std::find()

```
struct node {
    value_type value;
    struct node* next;
};

node* find(node* first, node* last,
     const value_type& value)
{
    for (; first != last;
         first=first->next)
        if (first->value == value)
            return first;
    return last;
}
```

Listing A.2: PULSE realization for std::find()

```
class list_find : chase_iterator {

    init(void *value, void* first) {
        *SP_PTR_VALUE = value;
        cur_ptr = first;
    }

    void* next() {
        return cur_ptr->next;
    }

    bool end() {
        if (*SP_PTR_VALUE ==
             cur_ptr->value) {
            *SP_PTR_RETURN = cur_ptr;
            return true;
        }
        if (cur_ptr->next == NULL) {
            *SP_PTR_RETURN =
                 KEY_NOT_FOUND;
            return true;
```

B.2 List data structure in Boost library

~~**Listing A.3:** Boost realization for bimap::find~~()

```
struct node {
    key_type key;
    struct node* next;
    value_type value;
};
void* find(const key_type& key, const
     hash_type& hash) const
{
    // The bucket start pointer can be
        pre-computed before offloading
    std::size_t buc =
        buckets.position(hash(key));
    node_ptr start = buckets.at(buc)
    for(node_ptr x = start; x != NULL; x
        = x->next){
        if(key == x->key){
            return x;
        }
    }
    return NULL;
}
```

~~**Listing A.4:** PULSE realization for bimap::find~~()

```
class bimap_find : chase_iterator {
public:
    key_type key;

    init(void *key, void* start) {
        *SP_PTR_KEY = key;
        cur_ptr = start;
    }

    void* next() {
        return cur_ptr->next;
    }

    bool end() {
```

B.3 Tree data structure in Google library

Listing A.5: Google realization for btree::internal_locate_plain_compare()

```
#define kNodeValues 8
struct btree_node {
   bool is_leaf;
   int num_keys;
   key_type keys[kNodeValues];
   btree_node* child[kNodeValues + 1];
};
IterType btree::internal_locate_plain_compare(const key_type &key, IterType iter) const {
   for (;;) {
      int i;
      for(int i = 0; i < iter->num_keys; i++) {
         if(key <= iter->keys[i]) {
            break;
         }
      }
      if (iter.node->is_leaf) {
         break;
      }
      iter.node = iter.node->child(i);
   }
   return iter;
}
```

Listing A.6: PULSE realization for btree::internal_locate_plain_compare()

```
class btree_find_unique : chase_iterator {
   init(void *key, void* iter) {
      *SP_PTR_KEY = key;
      cur_ptr = iter;
   }

```

B.4 Tree data structure in STL library

Listing A.7: C++ STL realization for map::find()

```
struct node {
    key_type key;
    node* left;
    node* right;
};

_M_lower_bound(node* x, node* y, const key_type& key)
{
    while (x != 0) {
        if (x->key <= key) {
            y = x;
            x = x->left;
        } else {
            x = x->right;
        }
    }
    return y;
}
```

Listing A.8: PULSE realization for map::find()

```
class map_find : chase_iterator {
    init(void *key, void* x, void* y) {
        *SP_PTR_KEY = key;
        *SP_PTR_Y = y;
        cur_ptr = x;
    }

    void* next() {
        if (cur_ptr->key <= *SP_PTR_KEY) {
            *SP_PTR_Y = cur_ptr;
            cur_ptr = cur_ptr->left;
        } else {
            cur_ptr= cur_ptr->right;
        }
        return cur_ptr->left;
    }
```

B.5 Tree data structure in Boost library

Listing A.9: Boost realization for avltree::find()

```
static node_ptr lower_bound_loop
(node_ptr x, node_ptr y, const KeyType
     &key)
{
    while(x){
        if(x->key >= key)) {
            x = x->right;
        }
        else{
            y = x;
            x = x->left;
        }
    }
    return y;
}
```

Listing A.10: PULSE realization for avltree::find()

```
class avltree_find : chase_iterator {
public:
    key_type key;
    void* y;

    init(void *key, void* x, void* y) {
        *SP_PTR_KEY = key;
        *SP_PTR_Y = y;
        cur_ptr = x;
    }

    void* next() {
        if(cur_ptr->key >= *SP_PTR_KEY) {
            cur_ptr = cur_ptr->right;
        }
        else{
            *SP_PTR_Y = cur_ptr;
            cur_ptr = cur_ptr->left;
```

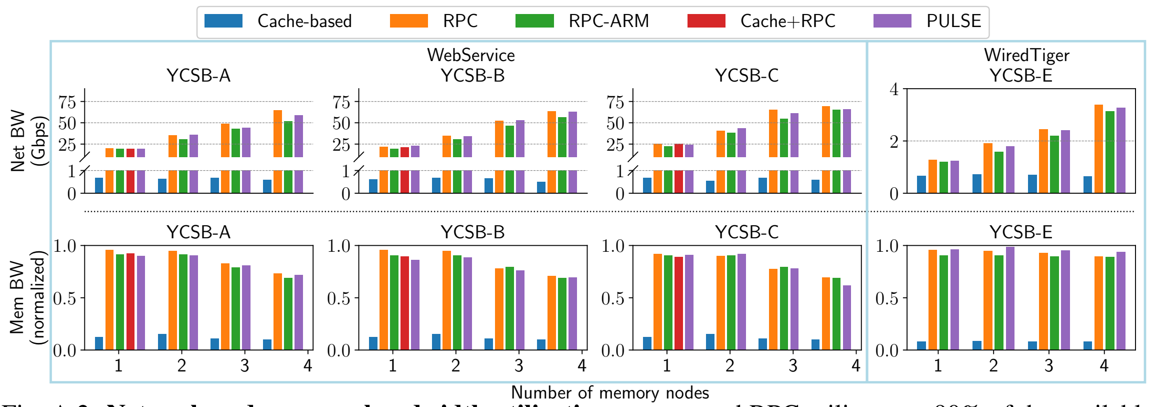


Fig. A.2: **Network and memory bandwidth utilization.** PULSE and RPC utilize over 90% of the available memory bandwidth, while the cache-based approach suffers from swap system overhead. In Webservice, the network bandwidth becomes the bottleneck due to large 8 KB data transfers.

C. PULSE Additional Evaluation Results

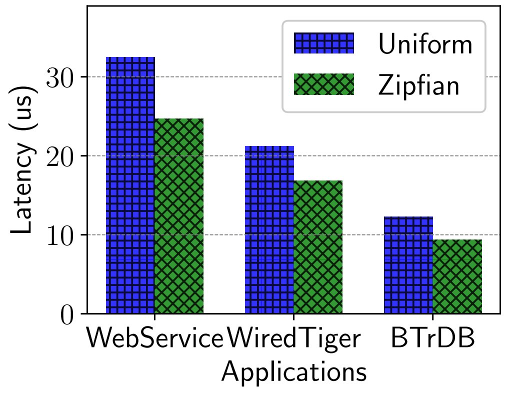
In this section, we provide additional evaluation results for PULSE.

C.1 Traditional Core Architecture vs. PULSE

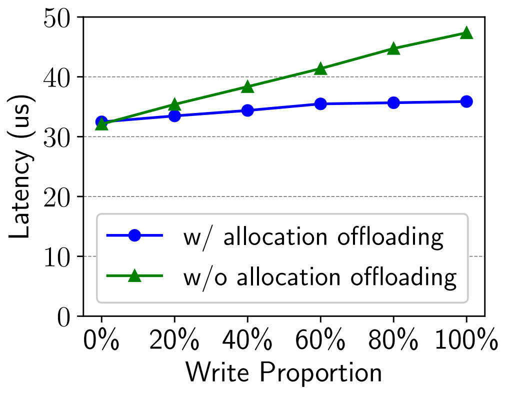
We evaluate the impact of the PULSE architectural design (§**??**) by comparing PULSE against PULSE-CORE, an in-order processor built on PULSE's components. We denote *C_x* as in tightly-coupled core architecture, where *x* is the number of cores. We denote *P_x_y* as PULSE architecture, where *x* is the number of logic pipelines and *y* is the number of memory pipelines. Table A.2 shows the power, performance, and area usage of various configurations. The performance metrics are obtained by executing the WebService application with various configurations. In PULSE's disaggregated architecture, when the number of logic and memory pipelines is equal to that of a traditional core architecture, power and area usage are higher due to additional logic and buffering in the interconnect and scheduler. However, due to the nature of pointer traversal operations (§**??**), PULSE requires fewer logic pipelines to achieve similar performance. For example, to fully saturate the memory bandwidth of a single node, PULSE uses only one logic pipeline and four memory pipelines, while a traditional core architecture requires four cores. As a result, PULSE saves 20.12% in power with only a 7.2% latency overhead, primarily due to the additional scheduler and data movement between workspaces.

Config	Pwr (W)	LUT %	BRAM %	Tpt (Mops/s)	Lat (us)
C_1	67.76	14.73	14.57	0.41	33.25
C_2	75.47	20.46	18.73	0.63	33.73
C_3	84.57	28.66	31.83	0.87	34.66
C_4	89.77	37.10	34.17	1.20	35.11
P_1_1	56.74	11.76	16.34	0.51	37.57
P_1_2	59.47	14.87	18.38	0.73	36.74
P_1_3	64.78	16.64	22.37	1.01	38.46
P_1_4	72.47	18.37	25.84	1.24	38.37
P_2_1	67.37	17.73	20.37	0.48	40.27
P_2_2	77.37	21.38	22.38	0.76	39.47
P_2_3	81.21	26.22	26.76	0.99	41.37
P_3_3	86.15	37.21	30.12	1.03	40.98
P_2_4	83.21	30.13	31.21	1.19	40.37
P_4_4	95.64	46.42	39.84	1.21	41.47

Table A.2: Comparison between traditional core architecture and PULSE architecture.

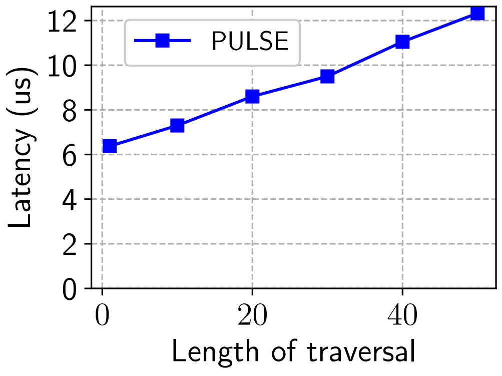


(a) Impact of access pattern

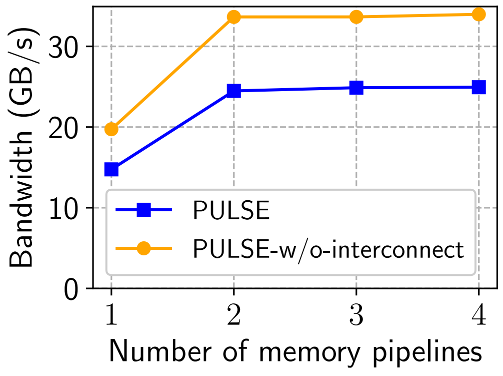


(b) Impact of modifications

Fig. A.3: (a) PULSE latency is up to 1.3× lower for skewed than uniform access patterns due to caching. (b) Offloaded allocations in PULSE improve the WebService request latencies as the proportion of writes increases by reducing the number of round trips per allocation.



(a) Traversal length



(b) Number of memory pipelines

Fig. A.4: **Sensitivity to traversal length and the number of memory pipelines.** (a) PULSE latency scales linearly with the length of traversal. (b) PULSE accelerator can saturate memory bandwidth with just two PULSE memory pipelines.

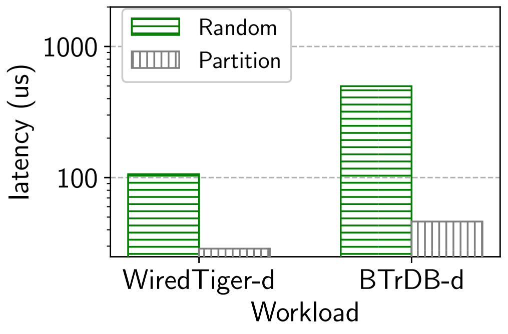
C.2 Network and Memory Bandwidth Utilization

We evaluate the network and memory bandwidth utilization of the three applications in Fig. A.2. For WiredTiger, PULSE and RPC utilize over 90% of the available memory bandwidth, while the Cache-based approach suffers from low network bandwidth and memory utilization due to swap system overhead. For WebService, the large 8 KB data transfers saturate the maximum bandwidth that the DPDK stack can sustain [125]. As a result, network bandwidth becomes the bottleneck, reducing PULSE and RPC memory bandwidth utilization under 3 and 4 memory nodes. The memory bandwidth is normalized, where 1.0 corresponds to 25 GB/s per node.

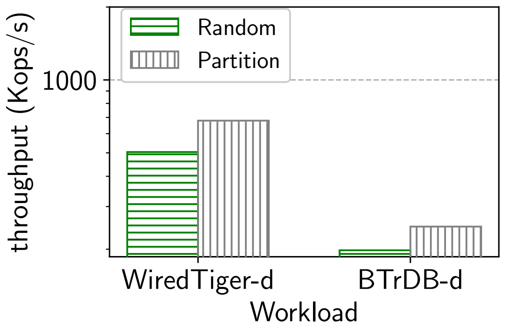
C.3 PULSE Sensitivity Analysis

We evaluate PULSE's sensitivity to workload characteristics and system parameters: access pattern, data structure modifications, traversal length, allocation policy, and the number of PULSE memory pipelines.

Impact of access pattern. While our evaluation so far has been confined to Zipfian workloads, we evaluate the impact of skewed access patterns on PULSE performance for all three applications. Our setup comprises a single 32GB memory node with a 2GB CPU node cache. Figure A.3(a) shows caching at the CPU node reduces the number of iterator requests offloaded to the PULSE accelerator for the skewed (Zipfian) workload, improving PULSE performance for such workloads by up to 1.33× relative to uniform ones.



(a) Latency



(b) Throughput

Fig. A.5: **Allocation policy.** PULSE performs better with the partitioned allocation since it minimizes cross-node traversals.

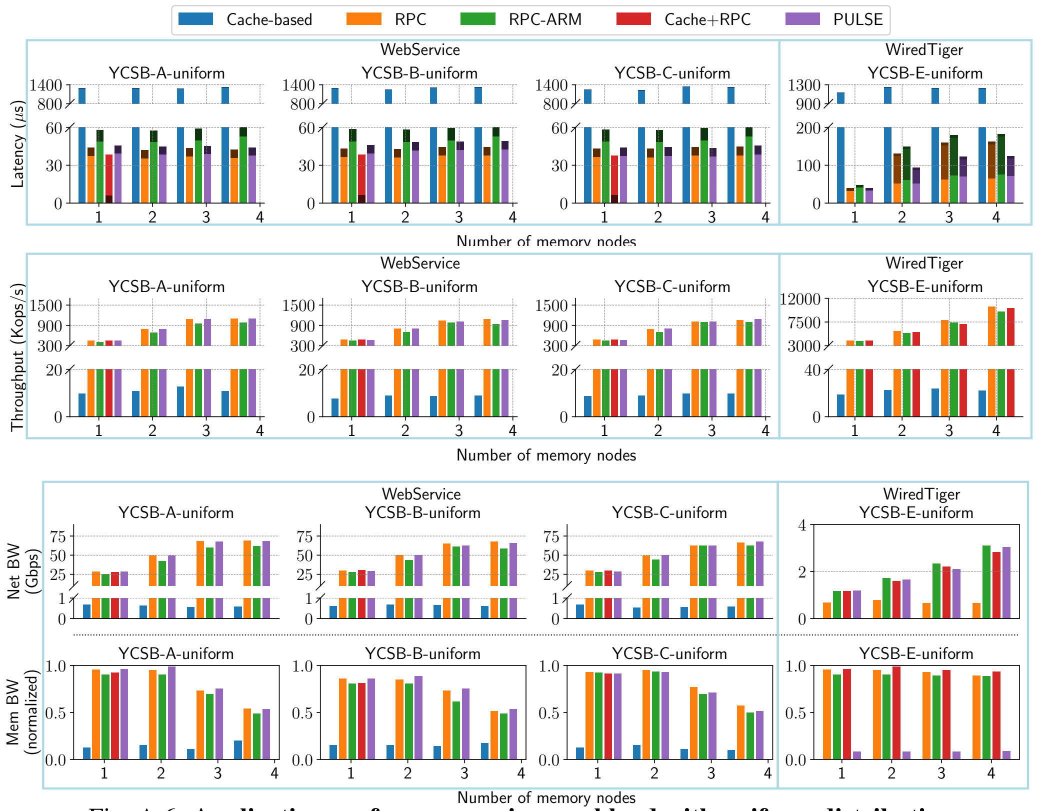


Fig. A.6: **Application performance using workload with uniform distribution.**

Impact of data structure modifications. Operations that modify data structures can require new memory allocations during traversal. Instead of returning control to the CPU node for allocations, PULSE populates the scratchpad for every request with a fixed number of pre-allocated memory regions. When a new allocation is initiated at the PULSE accelerator, it uses a pre-allocated memory region on the scratchpad. If all such regions (16 in our implementation) are used up in a single request, the traversal interrupts and returns to the CPU node. PULSE periodically replenishes pre-allocated entries, ensuring that allocation-triggered traversal interruptions are rare.

We evaluate the impact of data structure modifications in PULSE (§**??**) by increasing the proportion of writes for the WebService application on a single memory node. Figure A.3(b) shows that as the proportion of writes increases, PULSE without offloaded allocations experiences higher latencies (up to $1.4\times$) since each new node allocation requires two additional round trips; offloaded allocations reduce the allocation overhead to $< 1.1\%$.

Length of traversal. For simplicity, we evaluate traversal queries on a single linked list with varying numbers of nodes traversed per query. As expected, Fig. A.4(a) shows that the end-to-end execution latency for a linked list traversal scales linearly with the number of nodes traversed.

Allocation policy. We find that the allocation policy used for a data structure has a significant impact on application performance specifically for distributed traversals (Figs. A.5(a) and A.5(b)). We evaluated the WiredTiger and BTrDB workloads (that employ B+-Tree as their underlying data structure) with two allocation policies: one that partitions allocations in a way that ensures all nodes in half the subtree are placed on one memory node and the other half on another, and another that allocates memory uniformly across the two nodes (as in `glibc` allocator). The average latency for random allocations is $3.7-10.8\times$ higher than partitioned allocation since it incurs significantly more cross-node traversals. This shows that while uniformly distributed allocations can enable better system-wide resource utilization, it may be preferable to exploit application-specific partitioned allocations for workloads where performance is the primary concern.

Number of PULSE memory pipelines. We evaluate the number of PULSE memory access pipelines required to saturate PULSE's memory bandwidth on a single memory node. We used the same linked list as our traversal-length experiment due to its relatively low η value ($\sim$0.06), which allows us to stress the memory access pipeline without saturating the logic pipeline. Fig. A.4(b) shows that just 2 memory pipelines can saturate PULSE's the per-node memory bandwidth of 25 GB/s. We note that our 25 GB/s limit does not match the hardware-specified memory channel bandwidths; this is primarily due to our use of the vendor-supplied memory interconnect IP, required to connect all memory pipelines to all memory channels. Indeed, if we remove the IP and measure memory bandwidth when each memory pipeline is connected to a dedicated memory channel, PULSE can achieve a memory bandwidth up to 34 GB/s (shown as PULSE w/o Interconnect in Fig. A.4(b)).

PULSE performance with uniform workload. As illustrated in Fig. A.6, while sharing a similar trend as Zipfian distribution, all approaches experience higher latency compared to Zipfian distribution due to the ineffectiveness of caching. PULSE provides lower (vs. Cache-based, RPC-ARM, and Cache+RPC) or comparable (vs. RPC) latency for a single memory node and 2.2–29% lower latency (vs. RPC) for multi-memory nodes.

Bibliography

[1] S.-s. Lee, Y. Yu, Y. Tang, A. Khandelwal, L. Zhong, and A. Bhattacharjee. MIND: In-Network Memory Management for Disaggregated Data Centers. In *SOSP*, 2021.

[2] Y. Shan, Y. Huang, Y. Chen, and Y. Zhang. LegoOS: A Disseminated, Distributed OS for Hardware Resource Disaggregation. In *OSDI*, 2018.

[3] P. X. Gao, A. Narayan, S. Karandikar, J. Carreira, S. Han, R. Agarwal, S. Ratnasamy, and S. Shenker. Network Requirements for Resource Disaggregation. In *OSDI*, 2016.

[4] K. Asanović. FireBox: A Hardware Building Block for 2020 Warehouse-Scale Computers. 2014.

[5] S. Novakovic, A. Daglis, E. Bugnion, B. Falsafi, and B. Grot. Scale-out NUMA. In *ASPLOS*, 2014.

[6] L. Liu, W. Cao, S. Sahin, Q. Zhang, J. Bae, and Y. Wu. Memory Disaggregation: Research Problems and Opportunities. In *ICDCS*, 2019.

[7] K. Lim, J. Chang, T. Mudge, P. Ranganathan, S. K. Reinhardt, and T. F. Wenisch. Disaggregated Memory for Expansion and Sharing in Blade Servers. In *ISCA*, 2009.

[8] K. Lim, Y. Turner, J. R. Santos, A. AuYoung, J. Chang, P. Ranganathan, and T. F. Wenisch. System-level Implications of Disaggregated Memory. In *HPCA*, 2012.

[9] A. Samih, R. Wang, C. Maciocco, M. Kharbutli, and Y. Solihin. *Collaborative Memories in Clusters: Opportunities and Challenges*. 2014.

[10] Compute Express Link (CXL). https://www.computeexpresslink.org/.

[11] Y. Tang, P. Zhou, W. Zhang, H. Hu, Q. Yang, H. Xiang, T. Liu, J. Shan, R. Huang, C. Zhao, C. Chen, H. Zhang, F. Liu, S. Zhang, X. Ding, and J. Chen. Exploring performance and cost optimization with asic-based cxl memory. In *Proceedings of the Nineteenth European Conference on Computer Systems*, EuroSys '24, page 818–833, New York, NY, USA, 2024. Association for Computing Machinery.

[12] P. S. Rao and G. Porter. Is memory disaggregation feasible? a case study with spark sql. In *Proceedings of the 2016 Symposium on Architectures for Networking and Communications Systems*, ANCS '16, page 75–80, New York, NY, USA, 2016. Association for Computing Machinery.

[13] A. Verma, L. Pedrosa, M. Korupolu, D. Oppenheimer, E. Tune, and J. Wilkes. Large-scale cluster management at google with borg. In *Proceedings of the Tenth European Conference on Computer Systems*, EuroSys '15, New York, NY, USA, 2015. Association for Computing Machinery.

[14] C. Reiss, A. Tumanov, G. R. Ganger, R. H. Katz, and M. A. Kozuch. Heterogeneity and dynamicity of clouds at scale: Google trace analysis. In *Proceedings of the Third ACM Symposium on Cloud Computing*, SoCC '12, New York, NY, USA, 2012. Association for Computing Machinery.

[15] H. S. Gunawi, M. Hao, R. O. Suminto, A. Laksono, A. D. Satria, J. Adityatama, and K. J. Eliazar. Why does the cloud stop computing? lessons from hundreds of service outages. In *Proceedings of the Seventh ACM Symposium on Cloud Computing*, SoCC '16, page 1–16, New York, NY, USA, 2016. Association for Computing Machinery.

[16] T. Benson, A. Akella, and D. A. Maltz. Network traffic characteristics of data centers in the wild. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, page 267–280, New York, NY, USA, 2010. Association for Computing Machinery.

[17] M. R. Hines, A. Gordon, M. Silva, D. Da Silva, K. Ryu, and M. Ben-Yehuda. Applications know best: Performance-driven memory overcommit with ginkgo. In *2011 IEEE Third International Conference on Cloud Computing Technology and Science*, pages 130–137, 2011.

[18] D. Gupta, S. Lee, M. Vrable, S. Savage, A. C. Snoeren, G. Varghese, G. M. Voelker, and A. Vahdat. Difference engine: Harnessing memory redundancy in virtual machines. In *8th USENIX Symposium on Operating Systems Design and Implementation (OSDI 08)*, San Diego, CA, December 2008. USENIX Association.

[19] P. Bodík, I. Menache, M. Chowdhury, P. Mani, D. A. Maltz, and I. Stoica. Surviving failures in bandwidth-constrained datacenters. *SIGCOMM Comput. Commun. Rev.*, 42(4):431–442, aug 2012.

[20] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica. Dominant resource fairness: Fair allocation of multiple resource types. In *8th USENIX Symposium on Networked Systems Design and Implementation (NSDI 11)*, Boston, MA, March 2011. USENIX Association.

[21] R. Grandl, G. Ananthanarayanan, S. Kandula, S. Rao, and A. Akella. Multi-resource packing for cluster schedulers. *SIGCOMM Comput. Commun. Rev.*, 44(4):455–466, aug 2014.

[22] E. Amaro, S. Wang, A. Panda, and M. K. Aguilera. Logical memory pools: Flexible and local disaggregated memory. In *Proceedings of the 22nd ACM Workshop on Hot Topics in Networks*, pages 25–32, 2023.

[23] C. Hu, H. Huang, J. Hu, J. Xu, X. Chen, T. Xie, C. Wang, S. Wang, Y. Bao, N. Sun, et al. Memserve: Context caching for disaggregated llm serving with elastic memory pool. *arXiv preprint arXiv:2406.17565*, 2024.

[24] Redis. . https://redis.io/.

[25] J. Ousterhout, P. Agrawal, D. Erickson, C. Kozyrakis, J. Leverich, D. Mazières, S. Mitra, A. Narayanan, G. Parulkar, M. Rosenblum, et al. The Case for RAMClouds: Scalable High-performance Storage Entirely in DRAM. *SIGOPS OSR*, 2010.

[26] X. Zhang, U. Khanal, X. Zhao, and S. Ficklin. Making sense of performance in in-memory computing frameworks for scientific data analysis: A case study of the spark system. *J. Parallel Distrib. Comput.*, 120(C):369–382, oct 2018.

[27] A. Dragojević, D. Narayanan, O. Hodson, and M. Castro. FaRM: Fast Remote Memory. In *NSDI*, 2014.

[28] Z. Ruan, M. Schwarzkopf, M. K. Aguilera, and A. Belay. AIFM: High-Performance, Application-Integrated far memory. In *OSDI*, 2020.

[29] Q. Wang, Y. Lu, and J. Shu. Sherman: A write-optimized distributed b+tree index on disaggregated memory. In *SIGMOD*, 2022.

[30] P. Zuo, J. Sun, L. Yang, S. Zhang, and Y. Hua. One-sided {RDMA-Conscious} extendible hashing for disaggregated memory. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pages 15–29, 2021.

[31] E. Amaro, C. Branner-Augmon, Z. Luo, A. Ousterhout, M. K. Aguilera, A. Panda, S. Ratnasamy, and S. Shenker. Can Far Memory Improve Job Throughput? In *EuroSys*, 2020.

[32] J. Gu, Y. Lee, Y. Zhang, M. Chowdhury, and K. G. Shin. Efficient Memory Disaggregation with Infiniswap. In *NSDI*, 2017.

[33] I. Calciu, M. T. Imran, I. Puddu, S. Kashyap, H. A. Maruf, O. Mutlu, and A. Kolli. Rethinking software runtimes for disaggregated memory. In *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 79–92, 2021.

[34] C. Wang, H. Ma, S. Liu, Y. Li, Z. Ruan, K. Nguyen, M. D. Bond, R. Netravali, M. Kim, and G. H. Xu. Semeru: A {Memory-Disaggregated} managed runtime. In *14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20)*, pages 261–280, 2020.

[35] CHASE: Accelerating Distributed Pointer-Traversals on Disaggregated Memory. https://arxiv.org/pdf/2305.02388.pdf, 2023.

[36] A. Khandelwal, Y. Tang, R. Agarwal, A. Akella, and I. Stoica. Jiffy: Elastic far-memory for stateful serverless analytics. In *Proceedings of the Seventeenth European Conference on Computer Systems*, EuroSys '22, page 697–713, New York, NY, USA, 2022. Association for Computing Machinery.

[37] H. Li, D. S. Berger, S. Novakovic, L. R. Hsu, D. Ernst, P. Zardoshti, M. Shah, S. Rajadnya, S. Lee, I. Agarwal, M. D. Hill, M. Fontoura, and R. Bianchini. Pond: Cxl-based memory pooling systems for cloud platforms. *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2022.

[38] Y. Tang, P. Zhou, W. Zhang, H. Hu, Q. Yang, H. Xiang, T. Liu, J. Shan, R. Huang, C. Zhao, C. Chen, H. Zhang, F. Liu, S. Zhang, X. Ding, and J. Chen. Exploring performance and cost optimization with asic-based cxl memory. In *Proceedings of the Nineteenth European Conference on Computer Systems*, EuroSys '24, page 818–833, New York, NY, USA, 2024. Association for Computing Machinery.

[39] M. Vuppalapati, J. Miron, R. Agarwal, D. Truong, A. Motivala, and T. Cruanes. Building an elastic query engine on disaggregated storage. In *USENIX Networked Systems Design and Implementation (USENIX NSDI'20)*.

[40] M. Perron, R. Castro Fernandez, D. DeWitt, and S. Madden. Starling: A scalable query engine on cloud functions. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, page 131–141, New York, NY, USA, 2020. Association for Computing Machinery.

[41] Q. Pu, S. Venkataraman, and I. Stoica. Shuffling, fast and slow: scalable analytics on serverless infrastructure. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI'19)*, pages 193–206.

[42] A. Klimovic, Y. Wang, P. Stuedi, A. Trivedi, J. Pfefferle, and C. Kozyrakis. Pocket: Elastic ephemeral storage for serverless analytics. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 427–444, 2018.

[43] Y. Kim and J. Lin. Serverless data analytics with Flint. In *IEEE International Conference on Cloud Computing (CLOUD '18)*, pages 451–455. IEEE.

[44] Qubole Announces Apache Spark on AWS Lambda. https://www.qubole.com/blog/spark-on-aws-lambda.

[45] J. Carreira, P. Fonseca, A. Tumanov, A. Zhang, and R. Katz. Cirrus: A serverless framework for end-to-end ml workflows. In *Proceedings of the ACM Symposium on Cloud Computing*, SoCC '19, page 13–24, New York, NY, USA, 2019. Association for Computing Machinery.

[46] S. Fouladi, R. S. Wahby, B. Shacklett, K. V. Balasubramaniam, W. Zeng, R. Bhalerao, A. Sivaraman, G. Porter, and K. Winstein. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI'17)*, pages 363–376.

[47] E. Jonas, Q. Pu, S. Venkataraman, I. Stoica, and B. Recht. Occupy the cloud: distributed computing for the 99%. In *Proceedings of the 2017 Symposium on Cloud Computing*, pages 445–451. ACM, 2017.

[48] V. Shankar, K. Krauth, Q. Pu, E. Jonas, S. Venkataraman, I. Stoica, B. Recht, and J. Ragan-Kelley. numpywren: serverless linear algebra. *arXiv preprint arXiv:1810.09679*, 2018.

[49] S. Fouladi, F. Romero, D. Iter, Q. Li, S. Chatterjee, C. Kozyrakis, M. Zaharia, and K. Winstein. From laptop to lambda: Outsourcing everyday jobs to thousands of transient functional containers. In *USENIX Annual Technical Conference (USENIX ATC'19)*.

[50] Amazon. Amazon Athena. https://aws.amazon.com/athena.

[51] Amazon. Amazon Aurora Serverless. https://aws.amazon.com/rds/aurora/serverless.

[52] Azure. Azure SQL Data Warehouse. https://azure.microsoft.com/en-us/services/sql-data-warehouse.

[53] V. Sreekanti, C. W. X. C. Lin, J. M. Faleiro, J. E. Gonzalez, J. M. Hellerstein, and A. Tumanov. Cloudburst: Stateful functions-as-a-service. *arXiv preprint arXiv:2001.04592*, 2020.

[54] H. Zhang, Y. Tang, A. Khandelwal, J. Chen, and I. Stoica. Caerus: NIMBLE task scheduling for serverless analytics. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, pages 653–669. USENIX Association, April 2021.

[55] Q. Pu, S. Venkataraman, and I. Stoica. Shuffling, fast and slow: Scalable analytics on serverless infrastructure. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, pages 193–206, Boston, MA, February 2019. USENIX Association.

[56] M. Vuppalapati, J. Miron, R. Agarwal, D. Truong, A. Motivala, and T. Cruanes. Building an elastic query engine on disaggregated storage. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pages 449–462, Santa Clara, CA, February 2020. USENIX Association.

[57] K. Mahajan, M. Chowdhury, A. Akella, and S. Chawla. Dynamic query Re-Planning using QOOP. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 253–267, Carlsbad, CA, October 2018. USENIX Association.

[58] AWS Lamda. https://aws.amazon.com/lambda/.

[59] Google Cloud Functions. https://cloud.google.com/functions.

[60] Azure Functions. https://azure.microsoft.com/en-us/services/functions.

[61] Hadoop Distributed File System. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[62] C. Gray and D. Cheriton. *Leases: An efficient fault-tolerant mechanism for distributed file cache consistency*, volume 23. ACM, 1989.

[63] M. Burrows. The Chubby Lock Service for Loosely-coupled Distributed Systems. In *USENIX Conference on Operating Systems Design and Implementation (OSDI)*, pages 335–350, 2006.

[64] R. Droms. RFC 2131: Dynamic Host Configuration Protocol. https://www.ietf.org/rfc/rfc2131.txt, 1997.

[65] Amazon ElastiCache. https://aws.amazon.com/elasticache.

[66] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[67] Apache Spark. Unified engine for large-scale data analytics. https://spark.apache.org/.

[68] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. In *ACM SIGOPS operating systems review*, volume 41, pages 59–72. ACM, 2007.

[69] W. Lin, Z. Qian, J. Xu, S. Yang, J. Zhou, and L. Zhou. Streamscope: continuous reliable distributed processing of big data streams. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI'16)*, pages 439–453.

[70] R. Power and J. Li. Piccolo: Building Fast, Distributed Programs with Partitioned Tables. In *OSDI*, 2010.

[71] MongoDB: Sharded Cluster Balancer. https://docs.mongodb.com/manual/core/sharding-balancer-administration/#sharding-migration-thresholds.

[72] Ceph: Dynamic Bucket Index Resharding. https://docs.ceph.com/docs/mimic/radosgw/dynamicresharding/.

[73] Amazon Simple Notification Service (SNS). https://aws.amazon.com/sns.

[74] R. van Renesse and F. B. Schneider. Chain Replication for Supporting High Throughput and Availability. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 7–7, 2004.

[75] Apache Hadoop. https://hadoop.apache.org/.

[76] Amazon EC2. https://aws.amazon.com/ec2/.

[77] Amazon S3. https://aws.amazon.com/s3.

[78] Intel. *Barefoot Networks Unveils Tofino 2, the Next Generation of the World's First Fully P4-Programmable Network Switch ASICs*, 2018. https://bit.ly/3gmZkBG.

[79] EX9200 Programmable Network Switch - Juniper Networks. https://www.juniper.net/us/en/products-services/switching/ex-series/ex9200/.

[80] RoCE vs. iWARP Competitive Analysis. https://www.mellanox.com/related-docs/whitepapers/WP_RoCE_vs_iWARP.pdf, 2017.

[81] MySQL: Adaptive Hash Index. https://dev.mysql.com/doc/refman/8.0/en/innodb-adaptive-hash.html.

[82] SQLServer: Hash Indexes. https://docs.microsoft.com/en-us/sql/database-engine/hash-indexes?view=sql-server-2014.

[83] Teradata: Hash Indexes. https://docs.teradata.com/reader/RtERtp_2wVEQWNxcM3k88w/HmFinSvPP6cTIT6o9F8ZAg.

[84] R. Agarwal, A. Khandelwal, and I. Stoica. Succinct: Enabling Queries on Compressed Data. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2015.

[85] N. Askitis and R. Sinha. HAT-trie: A Cache-conscious Trie-based Data Structure for Strings. In *ACSC*, 2007.

[86] R. Bayer and E. McCreight. Organization and Maintenance of Large Ordered Indices. In *ACM-SIGMOD Workshop on Data Description, Access and Control*, 1970.

[87] A. Braginsky and E. Petrank. A lock-free b+tree. In *SPAA*, 2012.

[88] S. Heinz, J. Zobel, and H. E. Williams. Burst tries: a fast, efficient data structure for string keys. *TOIS*, 2002.

[89] A. Khandelwal, R. Agarwal, and I. Stoica. Blowfish: Dynamic storage-performance tradeoff in data stores. In *NSDI*, 2016.

[90] D. R. Morrison. PATRICIA - Practical Algorithm To Retrieve Information Coded in Alphanumeric. *JACM*, 1968.

[91] H. Zhang, H. Lim, V. Leis, D. G. Andersen, M. Kaminsky, K. Keeton, and A. Pavlo. Surf: Practical range query filtering with fast succinct tries. In *SIGMOD*, 2018.

[92] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin. PowerGraph: Distributed Graph-Parallel Computation on Natural Graphs. In *OSDI*, 2012.

[93] J. E. Gonzalez, R. S. Xin, A. Dave, D. Crankshaw, M. J. Franklin, and I. Stoica. GraphX: Graph Processing in a Distributed Dataflow Framework. In *OSDI*, 2014.

[94] A. Kyrola, G. E. Blelloch, and C. Guestrin. GraphChi: Large-Scale Graph Computation on Just a PC. In *OSDI*, 2012.

[95] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, 1999.

[96] L. Abraham, J. Allen, O. Barykin, V. Borkar, B. Chopra, C. Gerea, D. Merl, J. Metzler, D. Reiss, S. Subramanian, J. L. Wiener, and O. Zed. Scuba: Diving into data at facebook. *PVLDB*, 6(11), 2013.

[97] B. Berg, D. S. Berger, S. McAllister, I. Grosof, S. Gunasekar, J. Lu, M. Uhlar, J. Carrig, N. Beckmann, M. Harchol-Balter, and G. R. Ganger. The CacheLib caching engine: Design and experiences at scale. In *OSDI*, 2020.

[98] N. Bronson, Z. Amsden, G. Cabrera, P. Chakka, P. Dimov, H. Ding, J. Ferris, A. Giardullo, S. Kulkarni, H. Li, M. Marchukov, D. Petrov, L. Puzar, Y. J. Song, and V. Venkataramani. TAO: Facebook's distributed data store for the social graph. In *ATC*, 2013.

[99] R. Nishtala, H. Fugal, S. Grimm, M. Kwiatkowski, H. Lee, H. C. Li, R. McElroy, M. Paleczny, D. Peek, P. Saab, et al. Scaling memcache at facebook. In *NSDI*, 2013.

[100] X. Shi, S. Pruett, K. Doherty, J. Han, D. Petrov, J. Carrig, J. Hugg, and N. Bronson. Flight-Tracker: Consistency across Read-Optimized online stores at facebook. In *OSDI*, 2020.

[101] J. Yang, Y. Yue, and K. V. Rashmi. A large scale analysis of hundreds of in-memory cache clusters at twitter. In *OSDI*, 2020.

[102] S. Idreos, F. Groffen, N. Nes, S. Manegold, S. Mullender, and M. Kersten. Monetdb: Two decades of research in column-oriented database architectures. *IEEE Data Eng. Bull.*, 35, 01 2012.

[103] VoltDB. http://voltdb.com/downloads/datasheets_collateral/technical_overview.pdf.

[104] B. Fan, D. G. Andersen, and M. Kaminsky. Memc3: Compact and concurrent memcache with dumber caching and smarter hashing. In *Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation*, nsdi'13, page 371–384, USA, 2013. USENIX Association.

[105] X. Yu, G. Bezerra, A. Pavlo, S. Devadas, and M. Stonebraker. Staring into the abyss: An evaluation of concurrency control with one thousand cores. *Proceedings of the VLDB Endowment*, 8, 11 2014.

[106] MemCached. http://www.memcached.org.

[107] WiredTiger Storage Engine. https://www.mongodb.com/docs/manual/core/wiredtiger/.

[108] M. P. Andersen and D. E. Culler. BTrDB: Optimizing storage system design for timeseries processing. In *14th USENIX Conference on File and Storage Technologies (FAST 16)*, pages 39–52, Santa Clara, CA, February 2016. USENIX Association.

[109] S.-Y. Tsai, Y. Shan, and Y. Zhang. Disaggregating persistent memory and controlling them remotely: An exploration of passive disaggregated key-value stores. In *ATC*, 2020.

[110] Z. Yu, Y. Zhang, V. Bravermann, M. Chowdhury, and X. Jin. NetLock: Fast, Centralized Lock Management Using Programmable Switches. In *SIGCOMM*, 2009.

[111] Standard containers. https://cplusplus.com/reference/stl/.

[112] Boost library. https://www.boost.org/.

[113] Java iterator. https://www.w3schools.com/java/java_iterator.asp.

[114] C++ std::iterator. https://en.cppreference.com/w/cpp/iterator/iterator.

[115] The LLVM Compiler Infrastructure. https://llvm.org/.

[116] A. Rivitti, R. Bifulco, A. Tulumello, M. Bonola, and S. Pontarelli. ehdl: Turning ebpf/xdp programs into hardware designs for the nic. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, pages 208–223, 2023.

[117] DPDK. https://www.dpdk.org/.

[118] C. Lattner and V. Adve. Llvm: a compilation framework for lifelong program analysis & transformation. In *International Symposium on Code Generation and Optimization, 2004. CGO 2004.*, pages 75–86, 2004.

[119] LLVM's Analysis and Transform Passes. https://llvm.org/docs/Passes.html#introduction.

[120] K. Koukos, D. Black-Schaffer, V. Spiliopoulos, and S. Kaxiras. Towards more efficient execution: A decoupled access-execute approach. In *Proceedings of the 27th International ACM Conference on International Conference on Supercomputing*, ICS '13, page 253–262, New York, NY, USA, 2013. Association for Computing Machinery.

[121] J. Gandhi, V. Karakostas, F. Ayar, A. Cristal, M. D. Hill, K. S. McKinley, M. Nemirovsky, M. M. Swift, and O. S. Ünsal. Range translations for fast virtual memory. *IEEE Micro*, 36(3):118–126, 2016.

[122] Xilinx Content Addressable Memory (CAM). https://www.xilinx.com/products/intellectual-property/ef-di-cam.html.

[123] XUP Vitis Network Example (VNx). https://github.com/Xilinx/xup_vitis_network_example.

[124] AXI4 Protocol Burst size. https://bit.ly/3Bxh35b.

[125] A. Kalia, M. Kaminsky, and D. Andersen. Datacenter RPCs can be general and fast. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, pages 1–16, Boston, MA, February 2019. USENIX Association.

[126] Intel Xeon Gold 6240 Processor datasheet. https://ark.intel.com/content/www/us/en/ark/products/192443/intel-xeon-gold-6240-processor-24-75m-cache-2-60-ghz.html.

[127] Intel(R) RDT Software Package. https://github.com/intel/intel-cmt-cat.

[128] NIVIDIA MELLANOX BLUEFIELD-2. https://network.nvidia.com/files/doc-2020/pb-bluefield-2-smart-nic-eth.pdf.

[129] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears. Benchmarking Cloud Serving Systems with YCSB. In *Proceedings of the 1st ACM Symposium on Cloud Computing*, SoCC '10, page 143–154, New York, NY, USA, 2010. Association for Computing Machinery.

[130] WiredTiger storage engine. https://docs.mongodb.com/manual/core/wiredtiger/.

[131] Y. Zhong, H. Li, Y. J. Wu, I. Zarkadas, J. Tao, E. Mesterhazy, M. Makris, J. Yang, A. Tai, R. Stutsman, and A. Cidon. XRP: In-Kernel storage functions with eBPF. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*, pages 375–393, Carlsbad, CA, July 2022. USENIX Association.

[132] A Multi-Resolution Plotter that is compatible with BTrDB. https://github.com/BTrDB/mr-plotter.

[133] E. M. Stewart, A. Liao, and C. Roberts. Open μpmu: A real world reference distribution micro-phasor measurement unit data set for research and application development. 2016.

[134] A. Ousterhout, J. Fried, J. Behrens, A. Belay, and H. Balakrishnan. Shenango: Achieving high CPU efficiency for latency-sensitive datacenter workloads. In *NSDI*, pages 361–378, Boston, MA, February 2019. USENIX Association.

[135] Xilinx Runtime Library (XRT). https://www.xilinx.com/products/design-tools/vitis/xrt.html.

[136] Running Average Power Limit – RAPL. https://01.org/blogs/2014/running-average-power-limit-%E2%80%93-rapl.

[137] armv8registers. https://developer.arm.com/documentation/100095/0002/system-control/aarch64-register-summary/aarch64-performance-monitors-registers.

[138] Z. Guo, Y. Shan, X. Luo, Y. Huang, and Y. Zhang. Clio: A hardware-software co-designed disaggregated memory system. In *ASPLOS*, 2022.

[139] CZ120 memory expansion module. https://www.micron.com/products/memory/cxl-memory.

[140] I. Kuon and J. Rose. Measuring the gap between fpgas and asics. In *Proceedings of the 2006 ACM/SIGDA 14th International Symposium on Field Programmable Gate Arrays*, FPGA '06, page 21–30, New York, NY, USA, 2006. Association for Computing Machinery.

[141] J. Ahn, S. Hong, S. Yoo, O. Mutlu, and K. Choi. A scalable processing-in-memory accelerator for parallel graph processing. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, pages 105–117, 2015.

[142] H. Asghari-Moghaddam, Y. H. Son, J. H. Ahn, and N. S. Kim. Chameleon: Versatile and practical near-dram acceleration architecture for large memory systems. In *2016 49th annual IEEE/ACM international symposium on Microarchitecture (MICRO)*, pages 1–13. IEEE, 2016.

[143] G. Dai, T. Huang, Y. Chi, J. Zhao, G. Sun, Y. Liu, Y. Wang, Y. Xie, and H. Yang. Graphh: A processing-in-memory architecture for large-scale graph processing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 38(4):640–653, 2018.

[144] F. Schuiki, M. Schaffner, F. K. Gürkaynak, and L. Benini. A scalable near-memory architecture for training deep neural networks on large in-memory datasets. *IEEE Transactions on Computers*, 68(4):484–497, 2018.

[145] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun. Processing data where it makes sense: Enabling in-memory computation. *Microprocessors and Microsystems*, 67:28–41, 2019.

[146] E. Lockerman, A. Feldmann, M. Bakhshalipour, A. Stanescu, S. Gupta, D. Sanchez, and N. Beckmann. Livia: Data-centric computing throughout the memory hierarchy. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 417–433, 2020.

[147] F. Tu, Y. Wang, Z. Wu, L. Liang, Y. Ding, B. Kim, L. Liu, S. Wei, Y. Xie, and S. Yin. Redcim: Reconfigurable digital computing-in-memory processor with unified fp/int pipeline for cloud ai acceleration. *IEEE Journal of Solid-State Circuits*, 58(1):243–255, 2022.

[148] A. Devic, S. B. Rai, A. Sivasubramaniam, A. Akel, S. Eilert, and J. Eno. To PIM or not for emerging general purpose processing in DDR memory systems. In *ISCA*, pages 231–244, 2022.

[149] Z. Wang, J. Weng, S. Liu, and T. Nowatzki. Near-stream computing: General and transparent near-cache acceleration. In *HPCA*, pages 331–345, 2022.

[150] X. Xie, P. Gu, Y. Ding, D. Niu, H. Zheng, and Y. Xie. Mpu: Memory-centric simt processor via in-dram near-bank computing. *ACM Transactions on Architecture and Code Optimization*, 20(3):1–26, 2023.

[151] O. Mutlu, S. Ghose, J. Gómez-Luna, and R. Ausavarungnirun. A modern primer on processing in memory. In *Emerging Computing: From Devices to Systems: Looking Beyond Moore and Von Neumann*, pages 171–243. Springer, 2022.

[152] G. F. Oliveira, J. Gómez-Luna, S. Ghose, A. Boroumand, and O. Mutlu. Accelerating neural network inference with processing-in-dram: from the edge to the cloud. *IEEE Micro*, 42(6):25–38, 2022.

[153] C. Eckert, A. Subramaniyan, X. Wang, C. Augustine, R. Iyer, and R. Das. Eidetic: An in-memory matrix multiplication accelerator for neural networks. *IEEE Transactions on Computers*, 2022.

[154] P. Chi, S. Li, C. Xu, T. Zhang, J. Zhao, Y. Liu, Y. Wang, and Y. Xie. Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory. *ACM SIGARCH Computer Architecture News*, 44(3):27–39, 2016.

[155] V. Seshadri and O. Mutlu. Simple operations in memory to reduce data movement. In *Advances in Computers*, volume 106, pages 107–166. Elsevier, 2017.

[156] Y. Kwon, Y. Lee, and M. Rhu. TensorDIMM: A practical near-memory processing architecture for embeddings and tensor operations in deep learning. In *MICRO*, pages 740–753, 2019.

[157] A. Boroumand, S. Ghose, M. Patel, H. Hassan, B. Lucia, R. Ausavarungnirun, K. Hsieh, N. Hajinazar, K. T. Malladi, H. Zheng, and O. Mutlu. CoNDA: Efficient cache coherence support for near-Data accelerators. In *ISCA*, pages 629–642, 2019.

[158] B. Y. Cho, Y. Kwon, S. Lym, and M. Erez. Near data acceleration with concurrent host access. In *ISCA*, pages 818–831, 2020.

[159] L. Ke, U. Gupta, B. Y. Cho, D. Brooks, V. Chandra, U. Diril, A. Firoozshahian, K. Hazelwood, B. Jia, H.-H. S. Lee, M. Li, B. Maher, D. Mudigere, M. Naumov, M. Schatz, M. Smelyanskiy, X. Wang, B. Reagen, C.-J. Wu, M. Hempstead, and X. Zhang. RecNMP: Accelerating personalized recommendation with near-memory processing. In *ISCA*, pages 790–803, 2020.

[160] Z. Wang, J. Weng, J. Lowe-Power, J. Gaur, and T. Nowatzki. Stream floating: Enabling proactive and decentralized cache optimizations. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 640–653. IEEE, 2021.

[161] X. Xie, Z. Liang, P. Gu, A. Basak, L. Deng, L. Liang, X. Hu, and Y. Xie. Spacea: Sparse matrix vector multiplication on processing-in-memory accelerator. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 570–583. IEEE, 2021.

[162] L. Ke, X. Zhang, J. So, J.-G. Lee, S.-H. Kang, S. Lee, S. Han, Y. Cho, J. H. Kim, Y. Kwon, et al. Near-memory processing in action: Accelerating personalized recommendation with axdimm. *IEEE Micro*, 42(1):116–127, 2021.

[163] G. Singh, M. Alser, D. S. Cali, D. Diamantopoulos, J. Gómez-Luna, H. Corporaal, and O. Mutlu. Fpga-based near-memory acceleration of modern data-intensive applications. *IEEE Micro*, 41(4):39–48, 2021.

[164] A. Olgun, J. G. Luna, K. Kanellopoulos, B. Salami, H. Hassan, O. Ergin, and O. Mutlu. Pidram: A holistic end-to-end fpga-based framework for processing-in-dram. *ACM Transactions on Architecture and Code Optimization*, 20(1):1–31, 2022.

[165] G. Dai, Z. Zhu, T. Fu, C. Wei, B. Wang, X. Li, Y. Xie, H. Yang, and Y. Wang. Dimmining: pruning-efficient and parallel graph mining on near-memory-computing. In *Proceedings of the 49th Annual International Symposium on Computer Architecture*, pages 130–145, 2022.

[166] P. Gu, X. Xie, Y. Ding, G. Chen, W. Zhang, D. Niu, and Y. Xie. ipim: Programmable in-memory image processing accelerator using near-bank architecture. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pages 804–817. IEEE, 2020.

[167] J. Gómez-Luna, Y. Guo, S. Brocard, J. Legriel, R. Cimadomo, G. F. Oliveira, G. Singh, and O. Mutlu. Evaluating machine learning workloads on memory-centric computing systems. In *2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 35–49. IEEE, 2023.

[168] O. Kocberber, B. Grot, J. Picorel, B. Falsafi, K. Lim, and P. Ranganathan. Meet the walkers: Accelerating index traversals for in-memory databases. In *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-46, page 468–479, New York, NY, USA, 2013. Association for Computing Machinery.

[169] K. Hsieh, S. Khan, N. Vijaykumar, K. K. Chang, A. Boroumand, S. Ghose, and O. Mutlu. Accelerating pointer chasing in 3d-stacked memory: Challenges, mechanisms, evaluation. In *International Conference on Computer Design (ICCD)*, 2016.

[170] A. Bhardwaj, C. Kulkarni, and R. Stutsman. Adaptive placement for in-memory storage functions. In *ATC*, 2020.

[171] C. Kulkarni, S. Moore, M. Naqvi, T. Zhang, R. Ricci, and R. Stutsman. Splinter: Bare-Metal extensions for Multi-Tenant Low-Latency storage. In *OSDI*, 2018.

[172] J. You, J. Wu, X. Jin, and M. Chowdhury. Ship Compute or Ship Data? Why Not Both? In *NSDI*, pages 633–651, 2021.

[173] S. Novakovic, Y. Shan, A. Kolli, M. Cui, Y. Zhang, H. Eran, B. Pismenny, L. Liss, M. Wei, D. Tsafrir, and M. Aguilera. Storm: A Fast Transactional Dataplane for Remote Data Structures. In *SYSTOR*, page 97–108, 2019.

[174] Q. Zhang, X. Chen, S. Sankhe, Z. Zheng, K. Zhong, S. Angel, A. Chen, V. Liu, and B. T. Loo. Optimizing data-intensive systems in disaggregated data centers with TELEPORT. In *SIGMOD*, pages 1345–1359, 2022.

[175] D. Sidler, Z. Wang, M. Chiosa, A. Kulkarni, and G. Alonso. Strom: Smart remote memory. In *EuroSys*, 2020.

[176] J. Shen, P. Zuo, X. Luo, T. Yang, Y. Su, Y. Zhou, and M. R. Lyu. FUSEE: A fully Memory-Disaggregated Key-Value store. In *USENIX FAST*, 2023.

[177] P. Li, Y. Hua, P. Zuo, Z. Chen, and J. Sheng. ROLEX: A scalable RDMA-oriented learned Key-Value store for disaggregated memory systems. In *21st USENIX Conference on File and Storage Technologies (FAST 23)*, pages 99–114, Santa Clara, CA, February 2023. USENIX Association.

[178] H. An, F. Wang, D. Feng, X. Zou, Z. Liu, and J. Zhang. Marlin: A concurrent and write-optimized b+-tree index on disaggregated memory. In *Proceedings of the 52nd International Conference on Parallel Processing*, ICPP '23, page 695–704, New York, NY, USA, 2023. Association for Computing Machinery.

[179] X. Min, K. Lu, P. Liu, J. Wan, C. Xie, D. Wang, T. Yao, and H. Wu. Sephash: A write-optimized hash index on disaggregated memory via separate segment structure. *Proc. VLDB Endow.*, 17(5):1091–1104, 2024.

[180] J. Shen, P. Zuo, X. Luo, Y. Su, J. Gu, H. Feng, Y. Zhou, and M. R. Lyu. Ditto: An elastic and adaptive memory-disaggregated caching system. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 675–691, New York, NY, USA, 2023. Association for Computing Machinery.

[181] D. Korolija, T. Roscoe, and G. Alonso. Do OS abstractions make sense on FPGAs? In *OSDI*, 2020.

[182] E. Amaro, Z. Luo, A. Ousterhout, A. Krishnamurthy, A. Panda, S. Ratnasamy, and S. Shenker. Remote memory calls. In *Proceedings of the 19th ACM Workshop on Hot Topics in Networks*, pages 38–44, 2020.

[183] W. Reda, M. Canini, D. Kostić, and S. Peter. RDMA is turing complete, we just did not know it yet! In *NSDI*, 2022.

[184] Y. Sun, Y. Yuan, Z. Yu, R. Kuper, I. Jeong, R. Wang, and N. S. Kim. Demystifying cxl memory with genuine cxl-ready systems and devices, 2023.

[185] H. Li, D. S. Berger, S. Novakovic, L. Hsu, D. Ernst, P. Zardoshti, M. Shah, I. Agarwal, M. Hill, M. Fontoura, et al. First-generation memory disaggregation for cloud platforms. *arXiv preprint arXiv:2203.00241*, 2022.

[186] A. Li, S. L. Song, J. Chen, J. Li, X. Liu, N. R. Tallent, and K. J. Barker. Evaluating modern gpu interconnect: Pcie, nvlink, nv-sli, nvswitch and gpudirect. *IEEE Transactions on Parallel and Distributed Systems*, 31(1):94–110, 2019.

[187] Y. Sun, Y. Yuan, Z. Yu, R. Kuper, I. Jeong, R. Wang, and N. S. Kim. Demystifying cxl memory with genuine cxl-ready systems and devices, 2023.

[188] Y. Tang, P. Zhou, W. Zhang, H. Hu, Q. Yang, H. Xiang, T. Liu, J. Shan, R. Huang, C. Zhao, C. Chen, H. Zhang, F. Liu, S. Zhang, X. Ding, and J. Chen. Exploring performance and cost optimization with asic-based cxl memory. In *Proceedings of the Nineteenth European Conference on Computer Systems*, EuroSys '24, page 818–833, New York, NY, USA, 2024. Association for Computing Machinery.

[189] H. A. Maruf, H. Wang, A. Dhanotia, J. Weiner, N. Agarwal, P. Bhattacharya, C. Petersen, M. Chowdhury, S. Kanaujia, and P. Chauhan. Tpp: Transparent page placement for cxl-enabled tiered-memory. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, ASPLOS 2023, page 742–755, New York, NY, USA, 2023. Association for Computing Machinery.

[190] What Are PCIe 4.0 and 5.0? `https://www.intel.com/content/www/us/en/gaming/resources/what-is-pcie-4-and-why-does-it-matter.html`.

[191] D. D. Sharma, R. Blankenship, and D. S. Berger. An introduction to the compute express link (cxl) interconnect, 2023.

[192] Intel Corporation. Intel launches 4^{th} gen xeon scalable processors, max series cpus. https://www.intel.com/content/www/us/en/newsroom/news/.

[193] AMD Unveils Zen 4 CPU Roadmap: 96-Core 5nm Genoa in 2022, 128-Core Bergamo in 2023. https://wccftech.com/intel-clearwater-forest-e-core-xeon-cpus-up-to-288-cores-higher-ipc-more-cache/.

[194] Leo Memory Connectivity Platform for CXL 1.1 and 2.0. https://www.asteralabs.com/wp-content/uploads/2022/08/Astera_Labs_Leo_Aurora_Product_FINAL.pdf.

[195] Montage Technology. Cxl memory expander controller (mxc). https://www.montage-tech.com/MXC,accessedin2023.

[196] K. Kim, H. Kim, J. So, W. Lee, J. Im, S. Park, J. Cho, and H. Song. Smt: Software-defined memory tiering for heterogeneous computing systems with cxl memory expander. *IEEE Micro*, 43(2):20–29, 2023.

[197] D. Gouk, S. Lee, M. Kwon, and M. Jung. Direct access, High-Performance memory disaggregation with DirectCXL. In *2022 USENIX Annual Technical Conference (USENIX ATC 22)*, pages 287–294, Carlsbad, CA, July 2022. USENIX Association.

[198] M. Ahn, A. Chang, D. Lee, J. Gim, J. Kim, J. Jung, O. Rebholz, V. Pham, K. Malladi, and Y. S. Ki. Enabling cxl memory expansion for in-memory database management systems. In *Proceedings of the 18th International Workshop on Data Management on New Hardware*, DaMoN '22, New York, NY, USA, 2022. Association for Computing Machinery.

[199] Intel Corporation. Intel Agilex® 7 FPGA and SoC FPGA I-Series. https://www.intel.com/content/www/us/en/products/details/fpga/agilex/7/i-series.html.

[200] I. Kuon and J. Rose. Measuring the gap between fpgas and asics. In *Proceedings of the 2006 ACM/SIGDA 14th International Symposium on Field Programmable Gate Arrays*, FPGA '06, page 21–30, New York, NY, USA, 2006. Association for Computing Machinery.

[201] J. Weiner. [PATCH] mm: mempolicy: N:M interleave policy for tiered memory nodes. https://lore.kernel.org/linux-mm/YqD0%2FtzFwXvJ1gK6@cmpxchg.org/T/.

[202] NUMA balancing: optimize memory placement for memory tiering system. https://lore.kernel.org/linux-mm/20220221084529.1052339-1-ying.huang@intel.com/.

[203] Tiered-Memory: Hot page selection. https://lore.kernel.org/lkml/20220622083519.708236-2-ying.huang@intel.com/T/.

[204] Transparent Page Placement for Tiered-Memory. https://lore.kernel.org/all/cover.1637778851.git.hasanalmaruf@fb.com/.

[205] David L Mulnix. Intel® Xeon® Processor Scalable Family Technical Overview. https://www.intel.com/content/www/us/en/developer/articles/technical/xeon-processor-scalable-family-technical-overview.html.

[206] A. Cho, A. Saxena, M. Qureshi, and A. Daglis. A case for cxl-centric server processors, 2023.

[207] A. Cho and et al. A Case for CXL-Centric Server Processors. https://arxiv.org/abs/2305.05033.

[208] J. Yi, B. Dong, M. Dong, R. Tong, and H. Chen. MT$\hat{2}$: Memory bandwidth regulation on hybrid NVM/DRAM platforms. In *20th USENIX Conference on File and Storage Technologies (FAST 22)*, pages 199–216, Santa Clara, CA, February 2022. USENIX Association.

[209] Intel Corporation. Intel® Performance Counter Monitor (Intel® PCM). https://github.com/intel/pcm.

[210] Intel Corporation. Intel Unveils Future-Generation Xeon with Robust Performance and Efficiency Architectures. https://www.intel.com/content/www/us/en/newsroom/news/intel-unveils-future-generation-xeon.html.

[211] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears. Benchmarking cloud serving systems with ycsb. In *Proceedings of the 1st ACM Symposium on Cloud Computing*,

SoCC '10, page 143–154, New York, NY, USA, 2010. Association for Computing Machinery.

[212] Tecton.ai. Managing your Redis Cluster. https://docs.tecton.ai/docs/0.5/setting-up-tecton/setting-up-other-components/managing-your-redis-cluster.

[213] Google Cloud. Memory management best practices. https://cloud.google.com/memorystore/docs/redis/memory-management-best-practices.

[214] Redis enterprise. https://redis.io/docs/about/redis-enterprise/, 2023.

[215] Auto Tiering Extend Redis Enterprise databases beyond DRAM limits. https://redis.com/redis-enterprise/technology/auto-tiering/#:~:text=Redis%20Enterprise's%20auto%20tiering%20lets,compared%20to%20only%20DRAM%20deployments.

[216] C. Zou, H. Zhang, A. A. Chien, and Y. Seok Ki. Psacs: Highly-parallel shuffle accelerator on computational storage. In *2021 IEEE 39th International Conference on Computer Design (ICCD)*, pages 480–487, 2021.

[217] TPC-H is a Decision Support Benchmark. https://www.tpc.org/tpch/.

[218] Ice Lake SP: Overview and technical documentation. (n.d.). Intel. https://www.intel.com/content/www/us/en/products/platforms/details/ice-lake-sp.html.

[219] 4th Gen Intel Xeon Processor Scalable Family, sapphire rapids. (n.d.). Intel. https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html#gs.3m5uv2.

[220] McDowell, S. (2023, December 18). Intel launches 5th generation "Emerald Rapids" Xeon processors. Forbes. https://www.forbes.com/sites/stevemcdowell/2023/12/17/intel-launches-5th-generation-emerald-rapids-xeon-processors/.

[221] Kennedy, Patrick. "Intel Shows Granite Rapids and Sierra Forest Motherboards at OCP Summit 2023." ServeTheHome, 26 Oct. 2023,. www.servethehome.com/intel-shows-granite-rapids-and-sierra-forest-motherboards-at-ocp-summit-2023-qct-wistron.

[222] Mujtaba, H. (2023, December 1). Intel Clearwater Forest E-Core Only Xeon CPUs to offer up to 288 cores. https://wccftech.com/intel-clearwater-forest-e-core-xeon-cpus-up-to-288-cores-higher-ipc-more-cache/.

[223] S. Yi, D. Kondo, and A. Andrzejak. Reducing costs of spot instances via checkpointing in the amazon elastic compute cloud. In *2010 IEEE 3rd International Conference on Cloud Computing*, pages 236–243, 2010.

[224] Amazon EC2 M7a Instances. https://aws.amazon.com/ec2/instance-types/m7a/, 2023.

[225] Amazon EC2 M7i Instances. https://aws.amazon.com/ec2/instance-types/m7i/, 2023.

[226] Intel Shows Granite Rapids and Sierra Forest Motherboards at OCP Summit 2023. https://www.servethehome.com/intel-shows-granite-rapids-and-sierra-forest-motherboards-at-ocp-summit-2023-qct-wistron/.

[227] Elastic Compute Service, Volcano Engine, Bytedance. https://www.volcengine.com/product/ecs.

[228] G. W. D. Patel. GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. https://www.semianalysis.com/p/gpt-4-architecture-infrastructure, 2023.

[229] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention, 2023.

[230] Lightllm A Light and Fast Inference Service for LLM. https://github.com/ModelTC/lightllm.

[231] Julien Simon.Smaller is Better: Q8-Chat LLM is an Efficient Generative AI Experience on Intel® Xeon® Processors. https://www.intel.com/content/www/us/en/developer/articles/case-study/q8-chat-efficient-generative-ai-experience-xeon.html.

[232] R. Pope, S. Douglas, A. Chowdhery, J. Devlin, J. Bradbury, A. Levskaya, J. Heek, K. Xiao, S. Agrawal, and J. Dean. Efficiently scaling transformer inference, 2022.

[233] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA, 2023. Association for Computing Machinery.

[234] Alpaca: A Strong, Replicable Instruction-Following Model. https://crfm.stanford.edu/2023/03/13/alpaca.html.

[235] D. S. Berger, D. Ernst, H. Li, P. Zardoshti, M. Shah, S. Rajadnya, S. Lee, L. Hsu, I. Agarwal, M. D. Hill, and R. Bianchini. Design tradeoffs in cxl-based memory pools for public cloud platforms. *IEEE Micro*, 43(2):30–38, 2023.

[236] D. Blanchfield. The cloud native convergence: A new era of data-intensive applications. https://elnion.com/2023/06/05/the-cloud-native-convergence-a-new-era-of-data-intensive-applications/.

[237] A. Abulila, V. S. Mailthody, Z. Qureshi, J. Huang, N. S. Kim, J. Xiong, and W.-m. Hwu. Flatflash: Exploiting the byte-accessibility of ssds within a unified memory-storage hierarchy. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '19, page 971–985, New York, NY, USA, 2019. Association for Computing Machinery.

[238] S.-P. Yang, M. Kim, S. Nam, J. Park, J. yong Choi, E. H. Nam, E. Lee, S. Lee, and B. S. Kim. Overcoming the memory wall with CXL-Enabled SSDs. In *2023 USENIX Annual Technical Conference (USENIX ATC 23)*, pages 601–617, Boston, MA, July 2023. USENIX Association.

[239] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[240] L. Floridi and M. Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[241] K. Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

[242] A. Agrawal, A. Panwar, J. Mohan, N. Kwatra, B. S. Gulavani, and R. Ramjee. Sarathi: Efficient llm inference by piggybacking decodes with chunked prefills. *arXiv preprint arXiv:2308.16369*, 2023.

[243] Y. Zhong, S. Liu, J. Chen, J. Hu, Y. Zhu, X. Liu, X. Jin, and H. Zhang. DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 193–210, Santa Clara, CA, July 2024. USENIX Association.

[244] P. Patel, E. Choukse, C. Zhang, A. Shah, Í. Goiri, S. Maleki, and R. Bianchini. Splitwise: Efficient generative llm inference using phase splitting. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 118–132. IEEE, 2024.

[245] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA, 2023. Association for Computing Machinery.

[246] C. Hu, H. Huang, J. Hu, J. Xu, X. Chen, T. Xie, C. Wang, S. Wang, Y. Bao, N. Sun, and Y. Shan. Memserve: Context caching for disaggregated llm serving with elastic memory pool, 2024.

[247] X. Miao, G. Oliaro, Z. Zhang, X. Cheng, H. Jin, T. Chen, and Z. Jia. Towards efficient generative large language model serving: A survey from algorithms to systems. *arXiv preprint arXiv:2312.15234*, 2023.

[248] J. Yao, H. Li, Y. Liu, S. Ray, Y. Cheng, Q. Zhang, K. Du, S. Lu, and J. Jiang. Cacheblend: Fast large language model serving for rag with cached knowledge fusion, 2024.

[249] Y. Sheng, L. Zheng, B. Yuan, Z. Li, M. Ryabinin, B. Chen, P. Liang, C. Ré, I. Stoica, and C. Zhang. Flexgen: High-throughput generative inference of large language models with a

single gpu. In *International Conference on Machine Learning*, pages 31094–31116. PMLR, 2023.

[250] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[251] Y. Liu, H. Li, K. Du, J. Yao, Y. Cheng, Y. Huang, S. Lu, M. Maire, H. Hoffmann, A. Holtzman, et al. Cachegen: Fast context loading for language model applications. *arXiv preprint arXiv:2310.07240*, 2023.

[252] Y. Sun, Y. Yuan, Z. Yu, R. Kuper, C. Song, J. Huang, H. Ji, S. Agarwal, J. Lou, I. Jeong, R. Wang, J. H. Ahn, T. Xu, and N. S. Kim. Demystifying cxl memory with genuine cxl-ready systems and devices. In *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '23, page 105–121, New York, NY, USA, 2023. Association for Computing Machinery.

[253] P. Labs. GPT-fast Simple and efficient pytorch-native transformer text generation. `https://github.com/pytorch-labs/gpt-fast.git`, 2024.

[254] Y. Leviathan, M. Kalman, and Y. Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR, 2023.

[255] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, 2021.

[256] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.

[257] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[258] I. Corporation. Intel® Xeon® Platinum Processor. https://www.intel.com/content/www/us/en/products/details/processors/xeon/scalable/platinum.html, 2024.

[259] M. Arif, A. Maurya, and M. M. Rafique. Accelerating performance of gpu-based workloads using cxl. In *Proceedings of the 13th Workshop on AI and Scientific Computing at Scale Using Flexible Computing*, FlexScience '23, page 27–31, New York, NY, USA, 2023. Association for Computing Machinery.

[260] S. Sano, Y. Bando, K. Hiwada, H. Kajihara, T. Suzuki, Y. Nakanishi, D. Taki, A. Kaneko, and T. Shiozawa. Gpu graph processing on cxl-based microsecond-latency external memory. In *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, SC-W '23, page 962–972, New York, NY, USA, 2023. Association for Computing Machinery.

[261] D. Gouk, S. Kang, H. Bae, E. Ryu, S. Lee, D. Kim, J. Jang, and M. Jung. Breaking barriers: Expanding gpu memory with sub-two digit nanosecond latency cxl controller. In *Proceedings of the 16th ACM Workshop on Hot Topics in Storage and File Systems*, HotStorage '24, page 108–115, New York, NY, USA, 2024. Association for Computing Machinery.

[262] A. Li, S. L. Song, J. Chen, J. Li, X. Liu, N. R. Tallent, and K. J. Barker. Evaluating modern gpu interconnect: Pcie, nvlink, nv-sli, nvswitch and gpudirect. *IEEE Transactions on Parallel and Distributed Systems*, 31(1):94–110, 2019.

[263] NVIDIA. CUDA Runtime API - Memory Management Functions. https://docs.nvidia.com/cuda/cuda-runtime-api/group__CUDART__MEMORY.html, 2024.

[264] anon8231489123. ShareGPT Vicuna unfiltered dataset. https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered, 2024.

[265] Meta. Llama 2 70B: An MLPerf Inference Benchmark for Large Language Models. https://mlcommons.org/2024/03/mlperf-llama2-70b, 2024.