Abstract—As cloud workloads increasingly adopt the fault-tolerant Function-as-a-Service (FaaS) model, demand for improved performance has increased. Alas, the performance of FaaS applications is heavily bottlenecked by the remote object store in which FaaS objects are maintained. We identify that the upcoming CXL-based cache-coherent disaggregated memory is a promising technology for maintaining FaaS objects. Our analysis indicates that CXL’s low-latency, high-bandwidth access characteristics coupled with compute-side caching of objects, provides significant performance potential over an in-memory RDMA-based object store.

We observe however that CXL lacks the requisite level of fault-tolerance necessary to operate at an inter-server scale within the datacenter. Furthermore, its cache-line granular accesses impose inefficiencies for object-granular data store accesses.

We propose Apta, a CXL-based object-granular memory interface for maintaining FaaS objects. Apta’s key innovation is a novel fault-tolerant coherence protocol for keeping the cached objects consistent without compromising availability in the face of compute server failures. Our evaluation of Apta using 6 full FaaS application workflows (totaling 26 functions) indicates that it outperforms a state-of-the-art fault-tolerant object caching protocol on an RDMA-based system by 21–90% and an uncached CXL-based system by 15–42%.

I. INTRODUCTION

The Function-as-a-Service (FaaS) model is quickly becoming the de facto standard for cloud developers. In FaaS, applications are composed as workflows of stateless functions, and the cloud provider then orchestrates and schedules the functions dynamically on a fleet of compute servers.

The stateless nature of functions is good for availability, scalability, and elasticity, but it inevitably forces state to be maintained externally. Indeed, data stores such as Amazon S3 [7] are used to maintain state and pass input/output data between the stateless functions in the workflow. These data stores are the backbone of FaaS platforms.

Splitting state and compute, however, has an intrinsic data movement cost. Our analysis of FaaS functions from the FunctionBench [43] and SeBS [11] benchmark suites shows that on average 96% of the execution time per FaaS function is spent in retrieving data from the S3 object store. Replacing the S3 object store with a RDMA-based in-memory object store improves the situation somewhat – with 51% of execution time spent in retrieving data – but the problem persists. Communication overheads still limits performance.

Insight: FaaS objects on CXL disaggregated memory. We observe that upcoming CXL-based hardware disaggregated memory [9], [70] is a promising avenue for maintaining FaaS objects. CXL pools memory resources onto a logically centralized, physically distributed, highly-available memory server, and allows compute servers to perform load/store remote memory accesses in hardware over the network. The memory server, as shown in Fig. 1, is equipped with specialized hardware controllers for performing frequent data plane operations and minimalist low-power processors to handle rarer control plane operations [31], [48]. Since CXL allows for loads and stores to be handled in hardware like in a traditional NUMA machine [49], [50], [74], CXL-based disaggregated memory allows for significantly lower latency and higher bandwidth compared to high-performance RDMA-based remote memory.

Furthermore, the recently announced CXL 3.0 specification [10] allows the compute server caches to transparently cache data from a shared region in disaggregated memory, which matches well with the access patterns of a FaaS object store. Because FaaS functions typically share objects between them, object accesses exhibit significant locality and are amenable to caching. Therefore, such object caching and the use of a locality-aware scheduling policy (schedule functions where objects it uses are cached) has the potential for significantly reducing data movement.

Our analysis shows that a FaaS object store over a CXL-based disaggregated memory system with support for object-
granular accesses, coupled with a locality-aware scheduling policy can improve performance of the aforementioned FaaS functions by a significant 2.3× over the state-of-the-art RDMA-based object store. This is the performance opportunity Ápta targets.

**CXL provides consistency but forgoes availability.** To preserve flexibility and maximize throughput, a cloud system *dynamically* schedules a function on any available compute server. This dynamism combined with compute-side caching results in FaaS objects being replicated, and it is imperative that the replicas be kept consistent. Because compute servers can fail or be unresponsive in the datacenter, it is important that the consistency protocol remains *available* in the presence of such failures: i.e., the protocol should not block indefinitely if any of the servers fail. Alas the CXL 3.0 protocol [76] (which is a conventional protocol that enforces the Single-Writer-Multiple-Reader (SWMR) invariant [65]), while enforcing strong consistency, fundamentally blocks in the presence of server failures: if a server sharing an object fails, a write to that sharer from any other server could indefinitely block waiting for an acknowledgment from the failed sharer. Thus, this naive application of a traditional multi-processor coherence protocol (non-fault-tolerant) for distributed disaggregated memory leaves CXL systems vulnerable to system crashes.

**Severity of the problem:** Building system resiliency is an important problem as servers frequently fail or become unavailable in a datacenter environment. Google has observed that up to 25% of service-level disruptions are caused by machine-level failures [6]. A study of errors in even the highest reliability petascale supercomputers has shown that network link and server faults causing job failures occur with a mean time between failures (MTBF) of 160 hours [14], [40]. Consequently, fault tolerance is a key tenet of FaaS platforms. This is precisely why FaaS applications have already embraced failures via idempotent functions [5], [28]: if a function fails while executing (e.g., due to a compute server failure) the FaaS function can simply recover by re-executing. Therefore, it is imperative that the underlying CXL-based object store operates correctly in the presence of such server failures.

**Consistency & availability via fault-tolerant coherence.** We transform a strongly consistent SWMR-enforcing coherence protocol into a highly-available protocol in the presence of compute server failures. The idea consists of two simple steps: *lazy invalidation and coherence-aware scheduling*. In the first step we move the invalidations out of the critical path of the write so a writer is not blocked indefinitely when a server caching the sharer fails. But because invalidations are moved off the critical path of the write, there is a window of inconsistency where caches may hold stale values. In the second step we make a simple change to the FaaS scheduler [47] allowing it to schedule functions *only* on servers where there are no pending invalidations – thereby enforcing strong consistency as well as availability. Ápta’s method for transforming the non-fault-tolerant coherence protocol into a highly-available one can easily be applied to upcoming versions of CXL.

**Contributions.**

1) We make the case for a CXL-based object store for FaaS with object-granular reads/writes (Sec II). Our analysis using stand-alone FaaS functions indicates that such a design can provide a 69× performance improvement over the Amazon S3-based FaaS system, and a 2.3× improvement over a RDMA-based system. We observe, however, that such a system must remain fault-tolerant, which the existing CXL protocol specification falls short of.

2) We introduce Ápta (Fig. 1) – a CXL-based object store that allows compute server-side caching of objects without compromising consistency or availability (Sec. III). Ápta is tailored for object-granular accesses and defines a fault-tolerant inter-server cache-coherence protocol that, together with the FaaS scheduler, enforces strong consistency and provides high-availability in the presence of server failures. We have verified safety and liveness of the protocol in a model checker.

3) We evaluate the performance of Ápta (Sec. IV) using 6 full FaaS applications (total of 26 functions) and show that it provides 21–90% execution time speedup over a state-of-the-art fault-tolerant RDMA-based object store and 15–42% speedup over a reliable CXL-based object store without caching.

4) We observe that amongst all state-of-the-art high-performance remote memories and object stores that support caching (Table I), Ápta has the highest performance, and the highest availability in the presence of compute server failures.

<table>
<thead>
<tr>
<th>TABLE I_taxonomy of state-of-the-art proposals</th>
</tr>
</thead>
<tbody>
<tr>
<td>System</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>SF [10]</td>
</tr>
<tr>
<td>Clio-KV [31]</td>
</tr>
<tr>
<td>LegoOS [79]</td>
</tr>
<tr>
<td>Faa$t [77]</td>
</tr>
<tr>
<td>Symantec [31]</td>
</tr>
<tr>
<td>OFC [64]</td>
</tr>
<tr>
<td>LegoOS [79]</td>
</tr>
<tr>
<td>Cytos [50]</td>
</tr>
</tbody>
</table>

II. MOTIVATION & ANALYSES

In this section, we first demonstrate the compelling performance reason to migrate FaaS object store to a disaggregated memory system (abbreviated as DM). Next, we illustrate why DM, while providing improved performance, falls short of providing the level of fault-tolerance required for the FaaS paradigm. Finally, we highlight inefficiencies when existing cache line granularity DM is used to design an object store.

A. The performance potential of a DM-based object store

We compare the performance of FaaS functions from FunctionBench [43] and SeBS [11] benchmark suites1 using three

1We exclude micro benchmarks and network benchmarks that are non-deterministic and sensitive to external system delays.

2
different object stores: Amazon S3, RDMA-based, and DM-based object stores. The functions execute two basic operations on the object store: \( \text{obj} \leftarrow \text{get(objID)} \) at the beginning and \( \text{put(objID, obj)} \) at the end, where \( \text{objID} \) is an identifier for an object \( \text{obj} \). The computations in the middle of these functions are often unoptimized which hides the true bottlenecks in the system. We envision that high-performance frameworks such as Google TensorFlow [39] and Facebook PyTorch [38] will be adopted for FaaS in the future. We therefore ran the functions with Intel OneAPI [37] which applies vectorization, parallelization, cache blocking and other architecture specific optimizations.

**FaaS functions experience high communication overheads with Amazon S3:** When using the S3 object store, a \( \text{get} \) operation downloads the object from a remote S3 server into the compute server memory; post computation the \( \text{put} \) operation uploads an object from the compute server into a remote S3 server. We take the median of 100 executions accounting for cold function and tail latency effects [23], [86].

We observe that on an average 96% of execution time is spent in communicating data from/to the S3 object store (Fig. 2, all unstripped bars). This shows that the execution of FaaS functions in the cloud today is severely limited by the latency of accessing data from object stores. While S3 is based on disk based storage servers, it employs several optimizations like replication and sharding [7] to provide the best performance among today’s production object stores.

**In-memory object stores do not alleviate the communication overheads:** High-performance RDMA-based in-memory object stores completely bypass the remote CPU to read (write) objects directly from (into) the memory of the remote object server [12], [30], [63]. The \( \text{get} \) and \( \text{put} \) operations were modified to use one-sided RDMA verbs, which runs over an Infiniband network ( Mellanox ConnectX-3 NIC on PCIe-gen3 x16) [17], [58]. RDMA-based object stores are faster than traditional in-memory data stores that operate over general purpose ethernet networks like Redis [75], memcached [59] or Amazon ElastiCache [4]. However, even with such modern RDMA-based data store, on average 51% of execution time is still spent in communicating objects (Fig. 2, all striped bars).

**Overcoming RDMA’s Achilles heel:** The RDMA-based approach has several fundamental characteristics that limit performance – the use of software libraries like libibverbs and libmlx4, the need to perform two DMA data copy operations (at source and at destination, copying data to/from RNICs Memory Region) and managing the memory regions with software-initiated per-server static connection queue pairs. Several works have analyzed these and other drawbacks of RDMA [29], [31], [49]. DirectCXL [29] quantifies that even with the same underlying physical interconnect, RDMA’s irreducible overheads makes \( \text{get/put} \) operations 2.2\( x \) slower than CXL-based DM. DM is the new approach that chip manufacturers and cloud providers are investing in. DM overcomes the drawbacks of RDMA by allowing all data plane operations to be handled in hardware, thereby providing lower latency and higher bandwidth.

**DM reduces communication overheads:** The object is retrieved from a load/store semantic DM system. All standards for building such a DM system (GenZ [25], OpenCAPI [70]) have coalesced under the CXL umbrella due to their synergistic goals. Currently however, there exist only early prototypes: (i) OpenCAPI-based DM [74], providing RTT latencies of 950 ns and a bandwidth of 12.5 Gib/s; (ii) CXL-based DM [29], providing a lower RTT latency of 500 ns and a higher bandwidth. We pessimistically model the worst-case latency and bandwidth of OpenCAPI for our DM system. Our modeled DM system lowers latency by 3\( x \) and improves bandwidth by 10\( x \) over the RDMA system [49], [74].

With DM, the fraction of execution time spent for communication in FaaS functions reduces to 13% of the total time, on average. This translates to a large reduction in execution times of the functions. Fig. 3 shows that the DM-based object store is able to achieve a 59\( x \) geometric speedup over Amazon S3 and a 2\( x \) speedup over the RDMA object store.

**Caching - an additional benefit of DM:** A CXL DM system transparently caches object cache lines in the (on-chip SRAM or DRAM) hardware caches of the compute server, thereby being served at a lower latency compared to a remote memory server access. Such caching is extremely effective for FaaS applications which exhibit good object access locality [62].
This is because full FaaS applications, defined a state machine workflow of multiple individual functions (“function chains”), demonstrate known communication patterns like producer-consumer and broadcast within them [87]. This communication implies that successor functions can potentially access objects produced by any of its predecessors. When functions read objects from compute server caches in the DM system their execution time further speedups by 2%-100% (Fig. 3 DM+caching, assumes a DRAM cache of DDR4-like latency).

**Summary:** Our analysis indicates that maintaining FaaS objects in DM and the caching benefits it provides largely mitigates the key performance bottleneck of the FaaS paradigm.

B. The lack of fault-tolerance in current DM systems

FaaS object stores, such as Amazon S3, are designed to provide fault-tolerant operation for a failure-prone datacenter environment. Object `get` and `put` atomically read and write entire objects, with all or nothing semantics. A `get` is also guaranteed to read the value of the most recent `put`, therefore providing a strong consistency model known as linearizability [7]. This greatly simplifies things for a FaaS developer who can simply assume that a `get` would return the object written by the most recent put in the workflow.

**Enforcing strong consistency in the presence of caching.**

In the caching DM system, enforcing strong consistency for the FaaS execution environment can be challenging. For example, consider a simple workflow consisting of three functions: \( f_1 \rightarrow f_2 \rightarrow f_3 \), where \( f_1 \) and \( f_3 \) read object X, while \( f_2 \) writes to X. Further, let us assume that \( f_1 \) is assigned to server C1 while \( f_2 \) is assigned to C2. When \( f_1 \) executes on C1, it would cache the object in C1. When \( f_2 \) writes the object, it would render the value cached in C1 stale. Suppose the FaaS scheduler chooses to schedule \( f_3 \) on C1, \( f_3 \) would then read the stale value of X, violating strong consistency.

One way to enforce strong consistency in the presence of caching is to employ a cache coherence protocol. Conveniently, CXL 3.0 specifies an inter-server MESI-based coherence protocol [76], that enforces the SWMR invariant. In the above example, the write from \( f_2 \) would invalidate the cached copy of X in C1, ensuring that when \( f_3 \) is scheduled on C1, it will read the most recent value written by C2, and not the stale value.

**Whither Fault tolerance?** It is imperative that the aforementioned inter-server cache coherence protocol operates correctly even when compute servers fail or become unavailable. (In this work we assume that the DM server is kept highly-available using techniques such as replication [72], [79] and power redundancy.) Alas, traditional coherence protocols can block in the presence of such failures. Consider the same example where \( f_1 \) caches object X in server C1. When \( f_2 \) executing on C2, writes to X, the coherence protocol would send an invalidation to C1 which holds the object. Now, should C1 fail or become unreachable the write from \( f_2 \) would simply block, waiting for an acknowledgment, thereby rendering the system unavailable. Even if C1 does not fail but is simply slow to acknowledge (e.g., due to network congestion), the write from \( f_2 \) would be impacted, which can lead to high tail latency – a critical issue for FaaS platforms [85].

C. Inefficiencies of DM for object stores

Current DM system standards specify fixed fine-grain data access, caching and coherence mechanisms. However, object reads/writes typically have widely variable sizes, ranging from bytes to MBs [18], [62]. This causes two key inefficiencies. CXL enables compute servers to read cache lines from memory server while objects frequently span multiple cache lines. Hence, reading an object will incur multiple round trips to the DM, owing to limited MSHRs (miss status handling registers).

Second, CXL permits single cache line atomic write while a `put` must atomically write an object of multiple cache lines to the DM. This incurs additional latency for software write-ahead-logging i.e., undo/redo logs. Our analysis for the above benchmarks shows that a CXL object store will incur an average of 32% and 89% higher latency for `get` and `put` respectively, compared to an optimized object granular DM (evaluation methodology in Sec. IV).

**Summary:** Supporting compute server caching mandates a fault-tolerant coherence protocol that enforces strong consistency in the presence of compute server failures. CXL-based DM systems fail to provide this. Second, existing CXL cache line granular accesses are ill suited for FaaS object granular accesses.

III. Design

Âpta’s goal is to design a DM-based object store for FaaS applications (Sec. III-A) that provides fault-tolerant coherence (Sec. III-B) and optimum performance (Sec. III-C). To accomplish this, Âpta designs DM hardware controllers and modifies runtime software (Sec. III-D). Sec. III-E walks through the working of the entire Âpta system when executing real-world FaaS applications.

A. Setting the stage: Designing a DM-based object store

This section describes how Âpta leverages the features of a CXL 3.0 based DM system to construct an object store.

**(a1) Sharing objects between FaaS functions through DM**

**Extend shared memory IPC**

CXL 3.0 allows compute servers to access a shared memory region on the memory server. The compute server OS discovers and manages this CXL memory device as per UEFI/ACPI specifications [84] and exposes the DM address space as an extended CPU-less NUMA region [49], [50], [70].

In Âpta, FaaS functions execute as independent processes on compute servers. To access a shared object, the `get` and `put` operations map a DM memory region (containing the object) into their virtual memory using shared memory inter-process communication (shmem IPC) [77]. The shmem IPC API is enhanced to allow function processes on different compute servers to mount/access a shared memory region.

To illustrate, Fig. 4 shows the two functions \( f_1 \) and \( f_2 \) executing on server C1 and C2 respectively, sharing the object
X of size 50MB through the DM system. On each compute server, the shmem IPC segment, where X resides, is located in an extended NUMA physical address space (cPA - blue dashed regions in Fig. 4). Just as in CXL, the access controls and page tables for end-to-end address translation from compute server process virtual address (VA) to the memory server physical address (mPA) are initialized and setup by the OS. Once mapped, the object is accessed by the CPU (during the compute phase of the function) using load/store on cPA address. When these accesses miss in the LLC, the request is routed to the “home node” of the extended NUMA region (DM controller on the compute server). The DM controller uses the mapping to verify permissions and provides the memory server physical addresses (mPA) to be accessed.

(a2) Caching objects in compute server caches ► Defining a caching policy

Åpta introduces minimal changes to compute server caches, making them almost oblivious to disaggregation (in the spirit of CXL). The get operation, when mapping the shared object, sets the memory region of objects larger than size of the LLC as uncacheable using PAT or MTRR [71]. On an LLC miss, the cache line is read from DM and allocated in compute server caches. Similarly, objects are also write-allocated in the LLC. This policy allows retaining data in the LLC for any expected future reuse.

Importantly, the put operation immediately writes all modified cached lines through DM, making the caches effectively write-through. This policy allows tolerating compute server failures since a compute server LLC never holds the only copy of the object, and the memory server always holds a valid copy. The LLC silently evicts any of the DM cache lines which are in shared state i.e., the LLC does not issue a PutS coherence request to the directory. This saves interconnect network bandwidth and avoids LLCs having to evict entire objects if one of the object’s cache lines is evicted.

(b1) Keeping the cached objects on compute servers coherent ► Tailored coherence mechanism and protocol

“Simplicity is prerequisite for reliability” - Edsger Dijkstra

Simplified coherence: Recall that a FaaS function reads the object from the memory server and the compute server caches it in shared state; a put writes-through to the memory server and subsequently invalidates all sharers of the object in other compute servers. This eliminates the need for Modified or Exclusive states and reduces the inter-server protocol to two stable states - Shared and Invalid. This simplified coherence protocol, designed for the execution model of FaaS functions, hardens Åpta against compute server faults.

This Åpta protocol is layered hierarchically over and above intra-server coherence protocol. The intra-server coherence protocol is unchanged and regardless of this protocol Åpta enforces different policies in the inter-server protocol. This hierarchical organization allows Åpta to track sharers at compute server granularity (not individual caches within them). The Åpta protocol is employed for all requests from the compute server to the DM server.

Coarse granularity tracking: The use of DM in FaaS systems is restricted to sharing objects. Thus, it suffices for Åpta to use variable-sized object granularity tracking for the coherence protocol, as opposed to cache line level tracking in traditional chip level coherence protocols. In other words, we use a single state to encapsulate the state of all cache lines within the object. This is tracked using an object unique triplet of (objID, base mPA, size).

(b2) Provide high-availability while enforcing strong consistency ► Lazy invalidation of sharers with coherence-aware scheduling

Recall, to enforce strong consistency of the caches, a put completes only when all servers caching that object are
invalidated; therefore, put can block when servers fail. This
invariant of any conventional coherence protocol called Single-
writer-multiple-reader (SWMR) is enforced by synchronously
invalidating all sharers in the critical path of the put.

In Apta, the sharers are sent an invalidation message
asynchronously, i.e., the put is acknowledged immediately
without waiting for the sharers to be invalidated. The sharers
that are sent invalidations are tracked off the critical path until
they acknowledge the invalidation messages.

This lazy invalidation policy: (a) allows the write to be
acknowledged at lower latency thereby improving performance
and (b) more importantly, because writes need not wait for
sharers to be invalidated, there is no risk of writes being
blocked, thereby ensuring fault-tolerance.

Whither Consistency? Note, however, that this asynchronous
protocol described above could violate SWMR (and hence
linearizability). This is because at the instant the put is
acknowledged there may be cached copies in other servers
yet to be invalidated.

Lazy linearizability with scheduler support. Apta enforces
linearizability lazily using a combination of the coherence
protocol and the FaaS runtime scheduler. More specifically,
Apta never schedules function invocations on servers with
pending invalidations – the §Scheduling Correctness Criterion.
This correctness criterion ensures there is no risk of reading
any yet-to-be-invalidated stale objects present in the caches.
More precisely, we are now in a position to assert Lemma 1.

Lemma 1. The coherence protocol ensures that a get returns
the value of most recent put to that object.

Proof. Consider a get to object X. When the get is about
to be scheduled, there are either pending invalidations to X
or there are none. If there are no pending invalidations, there
are no stale values and hence the get will return the latest
value as per the original synchronous protocol. If there are one
or more pending invalidations, the scheduler ensures that the
function containing the get is not scheduled on those servers
with pending invalidations, and hence there is no risk of get
reading a stale value.

Thus, the Apta coherence protocol ensures that the caches
on the compute servers where functions execute are strongly
consistent. Meanwhile, caches in compute servers where functions
are not executing can be stale without affecting consistency.
Another benefit of Apta’s lazy invalidation protocol is
the ability to perform coherence actions at line-rate. This is
particularly important for processing packets in the data plane
on DPUs or SmartNICs in the network [21], [69].

C. Addressing the inefficiencies of DM

This section describes Apta’s optimization to adapt the DM
system for object idiosyncrasies.

(c1) Object-granular reads  Via bulk cache line loads
Recall, for each object load request that miss in the LLC, the
DM controller on the compute server issues a single cache
line read request over the interconnect making it inefficient
for objects spanning multiple cache lines.

CXL 3.0 [10] does provides fixed block request semantics
(2 or 4 contiguous cache lines) with the block request size to
be specified in advance. However, objects have more variable
block sizes and compute server LLCs cannot specify the block
size in advance as they operate oblivious of objects.

Apta builds on CXL 3.0 to provide variable sized, bulk
cache line requests. It bulk reads all the objects cache lines
into the compute server cache in one round trip to the memory
server (similar to [12]), providing the lowest possible latency
and maximizing the interconnect bandwidth utilization. This
process is illustrated in Fig. 5, left. The GET controller
(optimized DM controller) issues the LLC’s read request over
the interconnect. The memory server reads all cache lines
constituting the object from DRAM memory. It replies with all
these cache lines and squashes/ignores any immediate requests
for this object from that compute server. The GET controller
receives all the prefetched cache lines and inserts them into
the respective cache sets in the LLC. The LLC forwards the
demand miss cache lines to lower level caches and the CPU.

(c2) Object-atomic writes  Transactional atomic durability
Recall, CXL permits atomic writes of single cache lines which
forces a put to use software transactions (libpmemobj API)
to write an object of multiple cache lines atomically to the
DM system. These transactions use software logging (undo or
redo) which adds significant number of additional instructions
per transaction, hurting latency and throughput.

Apta provides hardware transactions for object atomic
writes to improve performance (similar to [32], [54]). The
hardware transaction ensures that when an object put is
executed, either the entire object is persisted4 or, in case of
failures, any partial writes are collectively discarded. If the
transaction succeeds, the memory server overwrites the new
version into the objects memory area. If the transaction fails,
it is retried, assuming the cause of the failure is transient.
If a retry threshold is exceeded, the exception is reported to

4Recall the memory server is usually kept highly-available and persistent
an external FaaS infrastructure system and the entire function execution is considered to have failed.

In the compute server, the PUT controller, co-located with the CPU, flags for persistence all the cache lines written by an object put. The controller orchestrates an atomic transaction, using a one-phase commit protocol with the memory server (Fig. 5, right). When the memory server issues a commit response, the PUT controller clears the persistence flags.

D. Realizing Ápta’s architecture

We now detail the memory server components - data plane controllers, control plane software and the interaction between them required to realize Ápta. We also describe in detail the coherence protocol sketched out in the previous section.

1) Micro-architecture of data-plane controllers: The data plane on the memory server is composed of a conventional memory controller and the Ápta controller. The Ápta controller is composed of four modular sub-controllers as shown in Fig. 1. This section details the micro-architecture of these controllers, each providing a certain functionality.

Object Serving sub-controller (OSC):

- **Function:** Serving objects (bulk cache lines) when the GET controller requests an object’s cache line.

The OSC translates the requested mPA to an object triplet. For this, OSC walks the object mapping data structure, populated by the FaaS runtime object manager (See III-D2). Similar to page tables, this translation latency can be reduced by using TLBs, page walk caches, cuckoo filters [82] etc. Once the physical address of the object is retrieved, the OSC issues memory access requests to the memory controller and replies to the compute server once it receives the data from it.

Object Persistence sub-controller (OPC):

- **Function:** Persists an entire object atomically into DM.

Recall, an object put initiates a one-phase commit protocol, between the PUT controller on the compute server and OPC on the memory server, to atomically write all the objects cache lines. As shown in Fig. 5 (right), first, the PUT controller on the compute server CPU sends a prepare message with the objID to be written. Then, it issues cacheline writeback (clwb [36]) for all the cache lines that are written by the put. The data from these cache lines, resident anywhere in the cache hierarchy of the compute server, are flushed to the memory server. OPC uses buffers (either using SRAM registers or a dedicated DRAM area) to temporarily stage cache lines written back from the compute server. OPC expects to receive a fixed number of cache line writes to complete the object write (inferred from the object triplet). Once it receives all cache lines of the object, it replies with a commit message, marking the end of transaction. OPC notifies object tracker controller (OTC) of the competition of an object write and flushes/drains the buffers to the memory controller.

Object Tracker sub-controller (OTC):

- **Function:** Directory for the Ápta coherence protocol.

Similar to a conventional directory, OTC maintains entries about the state and sharer vector for each object triplet. The OTC directory represents the ordering point for all requests. The directory is inclusive of all the compute server LLCs i.e., it holds directory entries for a superset of all objects cached in all the compute server LLCs. A miss in this directory cache indicates that the object is in state I. The sharer list tracked is not precise since the compute server LLC silently evicts blocks in shared state.

The protocol: OTC uses a simplified coherence protocol with Shared and Invalid stable states, avoiding the Modified and the Exclusive States. This is in line with the CXL specification as it flexibly allows implementations to use fewer stable states in the protocol. (We discuss further details of the CXL protocol in Sec. V-1.) Most coherence protocols involve transient states since transition from one stable state to another is not typically atomic [65]. Ápta re-purposes a transient state to account for the asynchronous invalidations. Fig. 6 illustrates the transition diagram for the OTC directory controller (with events and actions on stable and transient states). For ease of explanation, this protocol assumes that FaaS applications are race-free, i.e., no put or get can occur during an ongoing put. (However, the actual Ápta protocol can handle buggy FaaS applications with races as well.) A put event is triggered on the completion of an object persistence transaction by the OPC; a get event is triggered at the beginning of an object serving request by the OSC. An “Invalid” state for an object implies it is not cached in any compute server. A “Shared” state implies the object is cached in readable state in one or more compute servers. The transient state $S^A$ signifies that the new version of the object is cached in Shared state in one or more compute servers and there are pending invalidation-Acknowledgments from one or more compute servers for the old version of the object.

Suppose the directory receives a put for an object currently in shared state. Once the put transaction completes, the directory performs 3 actions in parallel - acknowledges the write, notifies the object invalidation controller (OIC) to send invalidation messages to all prior sharer compute servers, clears the old sharer vector and adds the compute server that requested the put as a sharer for the new version of the object (Recall the caching policy is write-allocate - compute server retains the object after a put). The directory then transitions to $S^A$ until it receives all invalidation-acknowledgments. While
in this state the directory can still service a get or put request for the object. For a get, the OSC responds with the latest object version, satisfying Lemma #1. Once OIC notifies that all outstanding invalidation-acknowledgments are received, the directory transitions back to shared state for the object.

The organization: The directory is organized as a standalone, set-associative directory cache structure at object granularity. When a directory cache set is full, the directory controller evicts a cold, shared object in the set. It issues an invalidate to all compute server sharers that cache this object and transitions to transient state SI. Any requests to the object while in this state, are stalled until all invalidations are acknowledged and an entry becomes available. Sizing the directory cache appropriately and correctly identifying cold objects can ensure that operations can continue at line-rate, without stalling.

Object Invalidation sub-controller (OIC):

> Function: Invalidates stale objects cached in compute servers.

The OTC requests the OIC to invalidate an object triplet (objID, base mPA, size) on a set of compute servers. OIC issues invalidation messages to the compute servers for all the object’s cache lines. At the compute server, the challenge however is translating the mPA address of the object to cPA address to issue invalidations to caches. This is achieved using an efficient object based reverse mapping [13], implemented in Linux for reserved mapping virtual memory (map chains)\(^5\). This reverse mapping is used by system calls like mmap, munmap, madvise etc. However, this is an expensive software call (measured to be ~1.4\(\mu\)sec per call) invoked by GET controller using interrupts and adds significant time overhead. Recall, in Åpta, invalidations are out of the critical path and hence this does not affect performance. Finally, the GET controller sends invalidation-acknowledgments.

The OIC tracks the number of invalidation-acknowledgments that are outstanding from each compute server using a counter. It notifies the OTC when all the invalidation-acknowledgments are received.

2) Control-plane software: Åpta modifies existing FaaS runtime control-plane software [47]. This software runs on the low-power SoC of the memory server. We outline the changes required in two of these components and describe their interface to the Åpta hardware controllers.

Executor Manager (EM):

> Function: Responsible for scheduling and tracking the execution of the state machine workflow of FaaS applications. EM selects a suitable compute server to schedule a function invocation and passes the invocation parameters to the function sandbox. EM scheduler is guided by the performance and correctness criteria when scheduling function invocations.

> Hardware interface: When scheduling functions, if the set of all objects to be accessed by the function is unknown (not declared), the scheduler queries the OIC to exclude all compute servers which have pending invalidation-acknowledgments. If the set of objects to be accessed by a function are declared in the state machine workflow, the scheduler looks up the object in OTC to determine where scheduling can be beneficial (current sharers) and if any compute servers are to be excluded (invalidation-acknowledgment pending).

Object Manager (OM):

> Function: Responsible for memory allocation and deallocation of objects in the memory server.

The objects are allocated in mPA in fixed bucket sizes (rounded up to the nearest fixed bucket size). The buckets are allocated as a contiguous physical memory address range, aligned at the cache line boundary in the memory server. This memory allocation strategy is akin to the memcached slab allocator [60] The OM runtime stores an object mapping data structure of mPA to unique objID, at a fixed location in memory. This data structure is organized as a radix tree followed by a trie\(^6\).

> Hardware interface: For serving objects, the OSC controller reads the object mapping data structure written by the OM runtime (from the fixed location in memory) and responds to the compute server mPA request with object-granularity bulk read semantics.

E. Putting it all together

Fig. 7 illustrates the application state machine workflow of three real world FaaS applications [46], [51] with the objects accessed by each function and annotated with an instance of scheduling decision made by the EM on a cluster of compute servers (C1 to C4) connected to access the Åpta object store.

We walk through the working of Åpta with the sentiment analysis application (Fig. 7, App 3) that evaluates customer reviews for products of a company and is triggered when the collated raw reviews file (csv) is uploaded to the object store.

When read.csv function on C1 receives an invocation trigger, the get call (rdata = get(“raw_data.csv”)) maps rdata to the shmem IPC region, located in the DM address range on C1. When rdata is accessed, the LLC miss triggers a request to the GET controller. The OSC controller responds with a set of cache lines of the object. All subsequent accesses to rdata in the computation hit in the caches. The put call atomically writes parsed_reviews object to memory server using hardware transaction between the PUT controller and OPC. C1 caches both raw_data.csv and parsed_reviews objects and accordingly, the OTC tracks C1 as a sharer of these objects.

Next, the sentiment_analysis function, scheduled on C2, similarly performs a get on parsed_reviews. On access the object is brought into the LLC, making C2 a sharer for the object. After computation, a put call writes a new version of parsed_reviews to the memory server. The Åpta

\(^5\)Originally an object in [13] referred to a memory mapped file which maps a range of data to a range of physical addresses. This works very well for our purposes since FaaS objects are also allocated contiguously within a range.

\(^6\)For object mapping, a combination of space efficient radix tree and lookup time efficient trie is used (inspired from page table in virtual memory and longest prefix match in routers, respectively). The radix tree traversal first points to 4KB/2MB page. Within the page, objects are organized as a trie. This organization ensures the data structure can be read in hardware controllers.
protocol acknowledges the write from C2 immediately and sends invalidation to C1 asynchronously, tracking C1 as having outstanding invalidation-acknowledgments. When scheduling the next set of parallel functions, the EM checks with OIC and does not schedule the functions on C1 to satisfy §Scheduling Correctness Criterion. Scheduling on C2 provides opportunity to exploit locality as it previously executed a predecessor function. Accordingly, publish_to_sns and write_to_db are scheduled on C2 and both functions benefit from cache hits for accesses to parsed_reviews.

IV. EVALUATING ÅPTA

A. Evaluation Goals:

(i) Compare performance of Åpta against the following state-of-the-art compute server fault-tolerant systems:

- RDMA-based object store with Faa$t caching [77]: An immutable object caching protocol run using two-sided RDMA verbs over Infiniband. On a put, write-through to object store with no sharer invalidation. On a get, if cache hit, incur one round trip to object store to ensure cached data is not stale (no object data transferred unless data is stale); if cache miss, read object from remote object store.
- faster RDMA-based object store with Faa$t caching: Uses the same interconnect as Åpta for RDMA, along with the above Faa$t software-based object caching protocol. This configuration allows us to isolate the benefits of improving just the underlying interconnect (transport layer).
- faster RDMA-based object store with Åpta caching: Uses the same interconnect and Åpta’s object caching protocol but in software. This configuration allows us to quantify the benefit of our optimized coherence protocol.
- CXL uncached DM: The cache line granularity DM of CXL, that achieves fault-tolerance by disabling caching of any DM data (requires no coherence protocol). This configuration allows us to quantify performance benefits of caching in DM and object semantic operations proposed for Åpta.

(ii) Demonstrate the fault-tolerance and resilience of Åpta
(iii) Break-down performance gains for get and put operations (compute time is kept constant for all configurations)
(iv) Evaluate robustness of performance gains with respect to varied interconnect properties and compute server capabilities.

B. Evaluation Methodology:

Our evaluation of Åpta is driven by a simulator based methodology (similar to DM proposals [8], [44], [45]). We now set out the configuration parameters and workloads used in the simulation of such a system.

Workloads: We use 6 full FaaS application workflows, totaling 26 functions, from different domains seen in FaaS - text, numeric, image, video processing. We simulate these full FaaS application workflows from start to finish to demonstrate realistic cache hit rate, invalidations and scheduling decisions. For each application, Table III shows the communication patterns in the workflow, input data size, constituent functions and a chosen instance of a schedule for an invocation.

For each application, Table III shows the communication patterns in the workflow, input data size, constituent functions and a chosen instance of a schedule for an invocation. These applications cover the full range of characterized input dataset/object sizes and function communication and invocation patterns [62], [66], [87]. The applications use local DRAM main memory to store intermediate data, akin to a scratchpad. Table III shows this measured local memory usage

![Fig. 7. FaaS applications annotated with object store interactions and scheduling decisions; highlight color changes indicate object writes requiring invalidation](image-url)
excluding the input object (max resident set size). We report the geometric mean speedup as an aggregate statistic across all applications.

> **System configuration:** We model a DM system with four servers (3 compute servers and 1 memory server). Each compute server has a single socket CPU with local DRAM memory. The CPU has per-core L1 and a socket-shared L2 cache, kept coherent with a directory-based MOESI protocol. Within the memory server, we simulate DDR4 DRAM memory, along with the ¯Apta controllers. The compute servers connect to the DM server with ordered point-to-point links of a fixed latency and bandwidth (full system config in Table II).

The RDMA configurations are measured on same hardware as in Sec. II. To model futuristic, faster RDMA (RDMAf - running over the same PCIe gen5 interconnect as ¯Apta), we add the latency overheads incurred for using RDMA operations and software coherence protocol to ¯Apta’s network latencies, as an approximation. Object get/put operations in a key-value store using RDMA to read/write data from/to remote memory are an average of 2.2× slower than CXL [29]. Above this, fault-tolerant coherent object get/put operations require employing key-value stores like FaaSt and Hermes [42], which use complex two-sided RDMA, adding even more latencies. We measured this as the latency difference between a write in Hermes and a one-sided RDMA write (≈14 μsec per call, fixed overhead irrespective of object size). We use these overhead latencies along with the respective coherence protocol actions to simulate RDMAf+FaaSt and RDMAf+¯Apta.

> **Simulator setup:** We simulate the identical shared memory version of the full FaaS applications written in python, compiled down to C. We generate traces of these programs using the Prism framework [67], which uses Valgrind to generates traces of compute, memory, thread create/join and barrier events. The tool produces synchronization and dependency-aware, architecture-agnostic traces. These traces are manually annotated with FaaS phases of execution i.e., get/compute/put.

We replay the traces in a modified gem5 simulator [78]. We implement the proposed inter-server ¯Apta coherence protocol and its hardware controllers. OSC and OTC lookup incur latency of 20 cycles each, modeled on average address translation and directory lookup latencies in modern processors.

![Fig. 8. Performance comparison of all configurations, normalized to RDMA without caching, for 6 full FaaS applications comprising of 26 FaaS functions](image)

### TABLE III

<table>
<thead>
<tr>
<th>Application (Patterns; Input data size; Max RSS)</th>
<th>Functions (compute server c1-c3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHI data [46] (Broadcast, Pipeline; 20KB; 100MB)</td>
<td>identityPHI (c1), delIdentity (c2), anonymize (c1), analytics (c1)</td>
</tr>
<tr>
<td>Sentiment Analysis [46] (Broadcast, Pipeline; 480KB; 91MB)</td>
<td>readcsv (c1), sentimentAnalysis (c2), publishSNS (c2), writeDB (c2)</td>
</tr>
<tr>
<td>FINRA [46] (Broadcast-Gather; 1.2MB; 23MB)</td>
<td>fetchMarket (c1), fetchPortfolios (c2), volume (c1), trade (c2), lasips (c3), side (c2), marginBalance (c1)</td>
</tr>
<tr>
<td>Video Transcode and Analysis [51] (Scatter-Gather, Pipeline; 2MB; 117MB)</td>
<td>locateKeyFrame (c1), splitVideo (c1), AnalyzeProcess (c1,c2), validate (c3), comact (c3)</td>
</tr>
<tr>
<td>Image Prediction [46] (Pipeline; 2.7MB; 357MB)</td>
<td>resize (c1), predict (c1), render (c1)</td>
</tr>
<tr>
<td>Serverless GEMM (sparse) [80] (Map-Reduce; 234KB; 943MB)</td>
<td>split (c1), mapper (c1,c2), split (c1), reducer (c1,c2)</td>
</tr>
</tbody>
</table>

C. Evaluation Results:

1) **Performance benefit and analysis:** Fig. 8 shows the performance of all configurations normalized to a baseline RDMA-based object store without caching.

**Result 1:** ¯Apta provides 42% geomean speedup over state-of-the-art RDMA+FaaSt. This performance gain comes from three sources: (a) improved network, (b) optimized ¯Apta coherence protocol, and (c) using hardware controllers for object access and coherence in DM. The RDMAf+FaaSt configuration provides 7% performance improvements over RDMA+FaaSt, showing the performance gains from just the improved network. Next, employing ¯Apta’s coherence protocol over RDMAf (RDMAf+¯Apta) provides further 12% improvement over the previous RDMAf+FaaSt, showing the performance gains from our optimized coherence protocol. Finally, ¯Apta’s use of DM hardware-controllers eliminates the irreducible software overheads of RDMAf, thereby providing 18% higher performance than previous RDMAf+¯Apta.

**Result 2:** ¯Apta provides 24% geomean performance gain over CXL-uncached by using a fault-tolerant object caching protocol and object semantic reads/writes. Note that employing the CXL-uncached object store will perform worse than a faster RDMA-based object store with caching (RDMAf+¯Apta), emphasizing the need for ¯Apta’s design in a DM system.

**Result 3:** We also evaluated the performance against the non fault-tolerant cached CXL DM. ¯Apta provides 10% geomean speedup over this CXL-cached system (not shown in graph). This shows that there is no performance cost to ¯Apta’s fault tolerance; in fact, ¯Apta shows a small improvement in

[33]. Memory ops are simulated with a detailed memory hierarchy. The replay mechanism uses FaaS phase of execution annotations to apply appropriate memory access characteristics for each phase of execution i.e., caches+local memory for compute, caches+DM for get/put. Integer and floating point ops are simulated with fixed CPI. We use an aggressive CPI and larger, lower latency L1/L2 caches to represent execution with optimized libraries (as in Sec. II-A). This simulator setup speeds up the computation phase of FaaS functions by 5× geomean compared to unoptimized (single thread) python functions run on a Intel i7-9700K machine. (Note this is conservative as Intel python extensions provide 200× speedup for scikit-learn, 90× for pandads, 3× for tensorflow [37].)
performance because it addresses CXL’s inefficiencies owing to its cache line granular accesses.

2) Fault-tolerance validation: We verify the complete Āpta protocol (with additional states to handle races, if any applications misbehave), in the Murϕ model checker [15] and exhaustively verified for liveness (deadlock-freedom) and safety (linearizability). Importantly, we also model check to prove correct and non-blocking behavior in the presence of sharer compute server failures. The detailed state transition table and Murϕ model for the protocol are available online at https://github.com/adarshpatil/apta.

Result 4: Because the Āpta protocol does not wait for acknowledgments in the critical path, it has the potential for lower tail latencies. To measure this, we run the applications 50 times under variable network latencies to reflect real world rack scale networks [73] and measure the standard deviation of execution times. The network requests experience a random latency within a Gaussian distribution (40% variation around the mean as measured for an Infiniband network [41]). On average, applications exhibit 32% lower standard deviation of execution time with Āpta compared to the non fault tolerant CXL-cached system, demonstrating the resilience of Āpta.

3) Performance Break-down: For the simulated schedule, Table IV shows the number of gets which hit in the cache (compulsory cache miss for first get request on all compute servers, while subsequent gets may potentially hit in the cache) and the number of puts that require sharer validations (these puts jeopardize DM system availability and increase latency with the blocking CXL-cached protocol).

Result 5: Āpta lowers geomean get latency by 90%, compared to RDMA+FaaS’s 57% reduction and CXL-uncached 71% reduction over baseline. Fig. 9, left shows the total get latency, normalized to baseline for each application. Although, both caching mechanisms (RDMA+FaaS and Āpta) see same cache hit rate, Āpta lowers geomean get latency by using an improved protocol and the DM interconnect.

Result 6: Āpta achieves the highest 81% reduction in geomean put latency, compared to 63% reduction for CXL-uncached. Fig. 9, right shows the baseline normalized total put latency. Since a put operation always writes through to the object store, RDMA and RDMA+FaaS see the same put latencies. Āpta achieves the reduction by using optimized hardware transactions over an improved DM interconnect.

4) Sensitivity studies: Āpta performance gain is subject to the compute-to-communication ratio of the application. Therefore, we study the performance sensitivity due to variations in interconnect characteristics and computation capabilities.

Result 7: The performance of Āpta improves with increase in bandwidth of the DM network, seeing 90% geomean speedup over baseline for 200Gbps. Interconnect latency has a smaller impact on the performance of Āpta. Āpta still provides a 84% geomean speedup with high latencies of 600ns, as expected for CXL switched fabrics. Fig. 10 a & b summarizes the speedups of Āpta over baseline, for varied interconnect network latencies and bandwidth.

Result 8: Āpta’s performance gain marginally reduces to 78% geomean with lower capability compute cores, as computation segment latency dominates in the total execution time. Fig. 10c shows the speedups obtained as we vary int/fp ops CPI of the core for both baseline and Āpta.

D. Evaluation Summary:
Āpta provides performance gains over all types of FaaS applications – from communication to compute heavy, applications with high object reuse and those with lower reuse, applications with serial and multiple parallel functions and over a range of object sizes.

V. DISCUSSION

1) Specifics of CXL support for Āpta: Āpta’s design introduces minimal changes to the CXL protocol and the servers. We discuss (i) the precise CXL protocol leveraged to design Āpta and (ii) changes needed to the CXL protocol specification to realize Āpta’s fault-tolerance benefit. We refer to relevant sections in the CXL 3.0 specification [10] in the discussion.

Using CXL_mem protocol for pooled shared memory: CXL 3.0 specification defines the creation of a pooled memory device where multiple compute servers are configured to access a single memory region concurrently - called “shared FAM”
VI. Related Work

Resilient coherence protocols: A class of works [1], [19], [20] design coherence protocols that can tolerate dropped and faulty messages. They reissue requests on a timeout to recover, but crucially assume all participants are alive. Ápta is the first work to handle complete node failures.

FaaS applications: A number of works [11], [43], [57], [85], [86], [90] composed function benchmarks and software stacks employed in FaaS platforms. They also demonstrate several FaaS inefficiencies: data communication, cold start etc. Ápta addresses a chief inefficiency of FaaS – data transfer overheads and provides a fault-tolerant DM system for FaaS applications.

Reducing communication overheads in FaaS: Several works [55], [64], [77], [81], [83], [88] aim to provide software-based caches at compute servers to cache objects. These works reinforce the potential of caching to improve performance despite being connected by fast networks. Ápta provides software transparent object caching using CXL-based DM.

In this paper, we have observed that upcoming CXL-based DM systems can alleviate the communication bottlenecks of cloud-based FaaS applications but lacks the necessary fault-tolerance to operate in a failure-prone datacenter. We have proposed Ápta, a CXL-based DM system for maintaining FaaS objects that provides efficient object-granular access and allows fault-tolerant caching of objects in compute servers caches, without compromising consistency. Thus, Ápta has showcased for the first time a fault-tolerant cloud use-case for CXL-based coherent disaggregated memory.


