

# Servo: Increasing the Scalability of Modifiable Virtual Environments Using Serverless Computing

Jesse Donkervliet  
Vrije Universiteit Amsterdam  
Amsterdam, Netherlands  
J.J.R.Donkervliet@vu.nl

Javier Ron  
KTH Royal Institute of Technology  
Stockholm, Sweden  
JavierRo@kth.se

Junyan Li  
Vrije Universiteit Amsterdam  
Amsterdam, Netherlands  
Junyan.Li@student.vu.nl

Tiberiu Iancu  
Vrije Universiteit Amsterdam  
Amsterdam, Netherlands  
T.Iancu@student.vu.nl

Cristina L. Abad  
ESPOL  
Guayaquil, Ecuador  
CAbad@fiec.espol.edu.ec

Alexandru Iosup  
Vrije Universiteit Amsterdam  
Amsterdam, Netherlands  
A.Iosup@vu.nl

**Abstract**—Online games with *modifiable virtual environments (MVEs)* have become highly popular over the past decade. Among them, Minecraft—supporting hundreds of millions of users—is the best-selling game of all time, and is increasingly offered as a service. Although Minecraft is architected as a distributed system, in production it achieves this scale by partitioning small groups of players over isolated game instances. From the approaches that can help other kinds of virtual worlds scale, none is designed to scale MVEs, which pose a unique challenge—a mix between the count and complexity of active in-game constructs, player-created in-game programs, and strict quality of service. Serverless computing emerged recently and focuses, among others, on service scalability. Thus, addressing this challenge, in this work we explore using serverless computing to improve MVE scalability. To this end, we design, prototype, and evaluate experimentally Servo, a serverless backend architecture for MVEs. We implement Servo as a prototype and evaluate it using real-world experiments on two commercial serverless platforms, of Amazon Web Services (AWS) and Microsoft Azure. Results offer strong support that our serverless MVE can significantly increase the number of supported players per instance without performance degradation, in our key experiment by 40 to 140 players per instance, which is a significant improvement over state-of-the-art commercial and open-source alternatives. We release Servo as open-source, on Github: <https://github.com/atlarge-research/opencraft>.

**Index Terms**—MVE, serverless, scalability, virtual environments, online games, backend, design, experimentation

## I. INTRODUCTION

Today’s digital societies increasingly leverage online games, which are cloud-deployed services architected as distributed systems, for other tasks with societal impact such as education [1], professional training [2], and social activism. Facilitating these use-cases requires support from a scalable *virtual world*. *Modifiable virtual environments (MVEs)* are a highly popular type of game which feature a virtual world that allows players to freely modify the game world. The canonical example of an MVE is Minecraft, which is the best-selling game of all time and supports over 140 million monthly active players [3]. Although MVEs are highly popular, their scalability in practice is very limited [4], with commercial services partitioning players into small, isolated game instances and

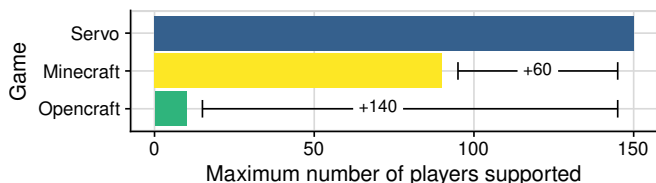


Fig. 1: The maximum number of supported players in Servo.

recommending as few as 10 (+1) maximum concurrent players per instance [5]. Existing techniques for scaling online games can help increase the number of players for a variety of game types, but do not address fundamental scalability challenges in MVEs. In contrast, in this work, we posit that MVE scalability can be addressed by leveraging serverless computing, and in particular, fine-grained deployment of applications as Function-as-a-Service (FaaS).

Online games have existed since at least the PLATO III distributed system [6, L.4,195] from the 1970s. Today, gaming is a lucrative industry that reaches a yearly revenue of \$175 billion [7]. An important part of the online gaming ecosystem are *modifiable virtual environments*, which are online games that feature a virtual environment that is modifiable, or even programmable, by players [8]. Prominent examples of MVEs are Minecraft, Roblox, and Second Life.

Fundamental scalability challenges in MVEs include numerous elements that do not appear in other types of games: its computationally-intensive virtual world containing in-game constructs that change state actively, player-created constructs that act as (repeated) programs, and other aspects we detail in Section II.

A promising technology to address MVE scalability challenges is serverless computing [9]–[11]. Serverless computing is a cloud-computing technology and another step towards providing computational resources as a utility. The canonical example of serverless computing is FaaS, which offers users serverless compute. However, serverless services are known to

exhibit high latency and latency variability, which makes them difficult to use in real-time systems such as online games.

In this work, we posit that serverless computing, and in particular, fine-grained deployment of applications as FaaS, can address MVE scalability when used in conjunction with latency-hiding mechanisms. Starting from this design principle, we design, implement, and conduct real-world experiments with Servo, a novel *serverless* backend architecture for scaling MVEs.

Figure 1 illustrates the real-world experimental results we obtain with a prototype of Servo, in comparison to Minecraft, the leading commercial MVE, and Opencraft, a state-of-the-art open-source MVE without a serverless architecture [12]. Servo scales to 150 players per instance, whereas Minecraft and Opencraft support up to 90 and 10 players, which is an increase of 60 and 140 players respectively.

*Existing approaches do not solve the challenge of scaling MVEs.* Figure 1 also illustrates that MVE workloads can pose significant stress to the gaming platform for even relatively few players, through other in-game elements that create complex workloads dynamically (which we detail in Section II-A). These elements raise very diverse scalability challenges that we aim to address in this work. Existing scalability techniques for online games [13]–[16], such as *zoning* and *replication*, can improve scalability for non-MVE games, but are not designed to take into account MVE scalability, which features complex in-game constructs that change state actively, player-created constructs that act as repeated programs, and other aspects we detail in Section II. As we explain in our technical report [17], zoning leads to frequent communication and coordination between game-servers, and replication even multiplies the workload caused by active in-game components.

*Scaling Minecraft-like MVEs can have high societal impact.* Through their modifiable environments, MVEs provide important benefits to society, providing entertainment, education, activism, and social interaction to an audience of more than 100 million people. With its large numbers of players and *mods*, MVEs provide world-wide entertainment, but they are also used for other beneficial societal needs. For example, MVEs are used for education in academia [18]–[20] and in industry with Microsoft’s Minecraft Education Edition [21]. MVEs are further used for social activism, with efforts including stopping illegal logging in Europe [22], and promoting press freedom around the globe [23]. Minecraft is not alone as a popular MVE. In fact, tens of thousands of other MVEs and MVE modifications (e.g., *mods*) exist [24], [25], many with large followings of millions of active players. Following this trend, Meta has already invested \$36 billion in developing an immersive, scalable MVE with advanced virtual-reality equipment for player-feedback [26]—a *metaverse* [27].

*Addressing the scalability challenges of MVEs and focusing on Minecraft-like games, we design, prototype, and evaluate Servo, a scalable, serverless, MVE system with fine-grained workload balancing. Our main Contributions are four-fold:*

**C1** We propose MVEs as a promising but challenging use case for serverless, and particularly, FaaS computing. We

describe this use case in Section II.

**C2** We design Servo, a serverless backend architecture for MVEs (in Section III). Ours is the first real-time online-multiplayer game using serverless design principles to improve the scalability of its virtual world. Although serverless simplifies building highly scalable and elastic applications, its programming model is not designed for online games, which operate under strict quality of service (QoS) requirements and need to continuously maintain consistency.

**C3** We conduct real-world experiments evaluating the performance of our Servo prototype (in Section IV). There is no standardized evaluation method for distributed systems, and benchmarking serverless systems is an active area of research [28], [29]. We design comprehensive real-world experiments that can help compare serverless and non-serverless MVEs, covering player-intensive workloads, workloads with complex in-game constructs, and other Minecraft-like workloads. From our key experiment on computational offloading (Section IV-B), we find that Servo can support up to 120 players during workloads with 200 in-game constructs, compared to 0 players(!) for Opencraft and Minecraft. For other workloads, Servo can increase the number of supported players under favorable conditions by 40 (+31%) to 140 (+1400%) players. These values summarize a complex suite of experiments, and not the exemplary result in Figure 1.

**C4** We implement a prototype on top of Opencraft, an open-source MVE and research platform. Our prototype runs on two popular commercial cloud platforms: Amazon Web Services (AWS) and Microsoft Azure. We release its source code on Github: <https://github.com/atlarge-research/opencraft>.

## II. MODIFIABLE VIRTUAL ENVIRONMENTS USE CASE

We argue in this section that modifiable virtual environments (MVEs) are an important but challenging use case for serverless computing. To this end, we propose and describe an operational model for MVEs, describe the performance characteristics of serverless computing, and describe the benefits and challenges of using serverless computing for online games.

### A. The Modifiable Virtual Environment

MVEs are online real-time interactive systems, or games, which allow players to explore virtual worlds. A main requirement of MVEs is to provide stable and good performance to players. To provide insight into how MVEs meet this requirement, we propose and describe here an operational model of MVEs, visible in Figure 2.

Users, or players, run an *MVE client* (①) on their personal device, which connects to an *MVE server* over the Internet. Players can only see and interact with other players connected to the same server. The main responsibilities of the client are to translate and relay user inputs to the server, and to show the user a visual representation of the virtual world by rendering

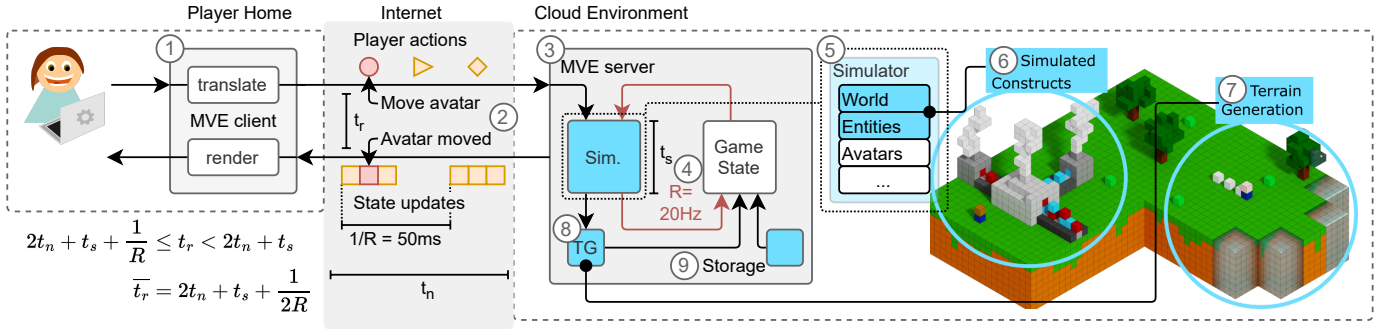


Fig. 2: An operational model of MVEs. Elements highlighted in blue indicate the components most relevant to this work. Whiskers indicate latency values. Components ⑥ and ⑦ are examples of MVE workloads absent in traditional online games.

state updates received from the server, typically at a fixed rate (e.g., 60 Hz).

Key to good performance is the game’s response time ( $t_r$ ), which is the time between a player performing an action on their client, and the result of this action becoming visible to all players connected to the same server. This metric is shown visually in Figure 2 using a “Move avatar” command as an example (②). Although *click-to-photon latency*, the time between a player clicking a button and the result appearing on the screen, is significantly affected by local and remote latency [30], [31], we focus in this work on the distributed systems aspects, modeling response time as dependent on the network latency ( $t_n$ ) and the performance of the server ( $t_s$ ).

Servers (③) are typically deployed in a cloud datacenter. Their main responsibility is to maintain a view of the global game state (for one instance of the virtual world) and simulate its state changes. The game simulates an immersive virtual world using its *simulator*, which takes as inputs player actions and the current game state, and produces state changes which are sent to players and stored internally to be used in the next iteration of the simulator.

The performance of the server depends on two main factors (④). First, the time it takes for the simulator (⑤) to complete one full iteration ( $t_s$ ). Second, the simulation rate ( $R$ ), i.e., the number of iterations per second. The simulation rate is typically fixed and varies per game type. For first-person perspective games,  $R$  is typically between 20 and 120, with Minecraft using an  $R$  of 20. Lowering the simulation rate increases the time budget for individual iterations, but also increases response time.

Providing good performance to users requires maintaining good response time, which requires maintaining the simulation rate,  $R$ , under varying workloads. This is challenging because online games, and especially MVEs, have highly dynamic workloads. These workloads exert monthly, weekly, and diurnal patterns, but also fluctuate moment-to-moment based on player behavior and environment simulation. For game service providers to offer these services cost-efficiently, this requires scalable and elastic resource management. Analyses of the temporal workload dynamics of online games are available

in existing work [32]. In this work, we focus on MVEs. Their modifiable, programmable, and procedurally generated terrain creates additional dynamic workloads for the server, which we describe below.

First, the game must process *simulated constructs* (SCs, ⑥). Players can modify the terrain by placing and removing blocks of varying types. Some block types, such as batteries and lamps, have internal state and can interact with other nearby blocks. For example, placing a battery next to a lamp may temporarily turn the lamp on. By connecting a collection of stateful blocks, players can *program* the MVE’s terrain. Throughout this article, we refer to such collection of stateful blocks as *simulated constructs* or SCs. Players can construct SCs of arbitrary size, within the limits imposed by the game, ranging from simple digital circuits to programmable digital computers, creating additional work for server instances, which must keep track of block states and their interactions with other blocks.

Second, procedural content generation (PCG, ⑦) is an important task in MVEs because players explore an infinite world that is generated on demand. When players move their avatar through the virtual world, the game must generate new parts of the world (i.e., terrain) with low latency, before the terrain data must be forwarded to the client.

Once a part of the terrain is generated, it is managed and persisted by the game server. Due to the world’s virtually infinite size, existing terrain must be loaded to and from persistent storage dynamically, based on avatar locations. Simulated constructs are only updated when their terrain is loaded in memory. When players leave the area, the terrain state is persisted (⑨) and its simulation is halted. When players approach areas that are not in memory, the MVE must retrieve these areas from storage, start updating any embedded simulated constructs, and forward the terrain data to clients. These actions must be completed before the terrain comes into the player’s view, and can increase the MVE’s bandwidth usage by an order of magnitude [4].

Although the generation and storage of terrain are not part of the game’s simulator, we identify two ways in which these elements can affect the game’s performance. First, a lack of

performance isolation can lead to interference, reducing the simulation rate when terrain generation is triggered. Second, even when the game meets its intended simulation rate, the separate generation component can fail to generate content at a sufficiently high rate, preventing players from moving to new areas and making the game unplayable.

To manage system complexity, our model makes two simplifying assumptions. First, it assumes network latency ( $t_n$ ) is symmetrical. Second, it assumes that state updates are only sent after the simulator has completed one full iteration. However, implementations can interleave the sending of state updates and simulating state changes.

### B. Serverless Computing

Serverless computing is an emerging computing service and a next step towards making digital services as simple to use as traditional utilities. Serverless computing promises effortless and cost-efficient scalability, but its characteristics make it challenging to adopt for certain application domains.

Although serverless computing does not have a generally-accepted precise definition, we follow in this work the definition from [33]. Following this definition, serverless computing: (1) Moves the responsibility of resource management and scheduling (RMS) from the end user to the cloud operator. Users merely provide application logic, and trust the cloud operator to scale the amount of resources to the current workload. (2) Employs fine-grained utilization-based billing. Users do not pay setup cost, or cost for idle resources, but only pay for resources used, with fine granularity (e.g., milliseconds).

Commercial clouds offer a plethora of serverless services. Well-known examples of these services include Function-as-a-Service (FaaS) (e.g., AWS Lambda and Azure Functions) and serverless storage (e.g., AWS S3 and Azure Blob Storage). With FaaS, users upload code to their cloud provider of choice and configure one or several events that trigger the execution of this code. When such an event occurs, the cloud provider schedules the execution of the user-provided code. This model has several appealing properties. First, it allows for simple but powerful scalability because every request is handled by its own instance. Second, the user only pays for the amount of time and resources consumed while executing their provided code, which means the user does not pay for setup costs nor gets charged when there are no requests.

Serverless storage allows users to read and write (binary) data to persistent storage in the cloud. Depending on the service, users are charged for the amount of data stored, the number of read and write operations, the number of bytes transferred over the network, or a combination of these metrics. In return, the cloud provider performs all necessary RMS, leaving the user with a simple read-write API. Commercial clouds offer a range of serverless storage services that make different performance-cost trade-offs. For example, services catering to long-term data archiving are generally cheaper to use, but have significantly higher read and write latencies, sometimes in the order of days.

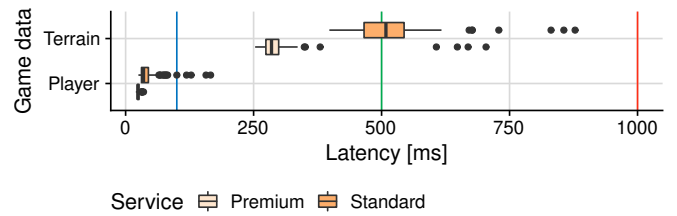


Fig. 3: Download latency (variability) from Azure Blob Storage for two types of game data. Vertical bars indicate approximate network latency thresholds for FPS (blue, left), RPG (green, middle) and RTS (red, right) games [35].

### C. Leveraging Serverless for MVEs

In this section, we describe the main benefits and challenges of using serverless computing for online real-time games generally, and MVEs specifically.

**Benefits:** First, games can benefit from the simple elasticity serverless computing provides because gaming workloads change significantly over time [34]. Furthermore, MVE workloads can increase or decrease significantly in the order of seconds, depending on player actions. For example, if a player builds and enables simulated constructs, a single player can exceed the resource requirements of a game instance (see Section IV). Second, fine-grained billing allows this elasticity to be cost-efficient and allows smaller development teams to participate in the market without large up-front investments.

**Challenges:** First, whereas games typically have stringent latency requirements in the order of tens of milliseconds, serverless services commonly exhibit high latency and significant performance variability, for example due to *cold starts*. Not meeting the game’s latency requirements can result in significant (visual) glitches for the player that can cause them to quit the game, causing the game developer to lose users and revenue.

An example of this phenomenon is shown in Figure 3. The figure shows the end-to-end latency of retrieving terrain-data and player-data from Azure Blob Storage. The former operation is performed when players move their avatars towards a previously unloaded area of the virtual world, which occurs frequently during play, whereas the latter is performed every time a player connects to a game instance, and their associated data is retrieved from persistent storage. For both storage plans, the latency variability is significant, as indicated by the width and length of the box and whiskers, and the value of the outliers. The vertical lines show approximate values for the maximum acceptable network latency for varying game genres [35]. Games with an omnipresent model (e.g., RTS games, in red), third-person games (e.g., RPGs, in green), and first-person games (e.g., FPSs, in blue). Most MVEs, including Minecraft, are first-person games, putting them in the games with the most restrictive network latency requirement (100 ms). This indicates that a naive application of serverless technology can easily break its performance requirements.

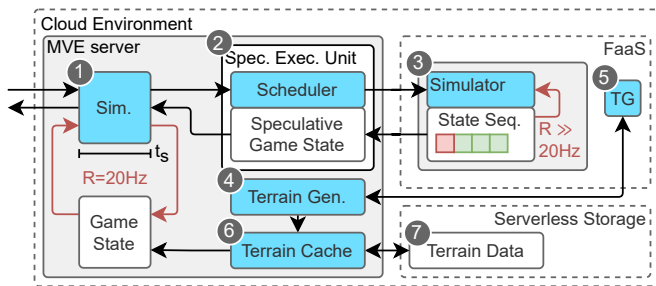


Fig. 4: Servo design overview. Simulated constructs (SC) and components ①, ②, etc. are discussed in the main text.

Second, FaaS services typically impose strict limits on the amount of time a function is allowed to run, whereas online games typically consist of long-running player sessions that can last several hours. Third, and finally, FaaS services typically do not allow external services to initiate communication to a running function instance, whereas games and other interactive systems require continuous communication between the system components to maintain a consistent state.

### III. SYSTEM DESIGN

In this section we discuss our design for Servo, a serverless backend architecture for MVEs which uses a collection of serverless techniques to provide fine-grained virtual-world scalability. We describe the requirements for such a system and formulate a novel design.

#### A. System Requirements

- R1 Independently scale MVE server components.** Although replicating monolithic MVE instances is the simplest approach to scale out, it is not efficient. Because players increase the workload on specific components of the MVE (e.g., terrain simulation, terrain generation), scaling these components independently limits wasted resources.
- R2 Simulation latency should not exceed 50 ms.** To provide an immersive experience to players, the game should maintain a constant simulation rate,  $R = 20$  Hz. This implies the simulation latency should not exceed  $1/R = 50$  ms (see Section II).
- R3 Limit additional resource management complexity for game operators.** An important benefit of the client-server architecture discussed in Section 2 is its simplicity: a game operator runs one server instance for every virtual world instance. A serverless game design should limit the additional number of steps game operators must perform.
- R4 System is fully transparent to users.** To improve usability and potential for community adoption, using the system should require zero effort from, and be fully transparent to, the users (i.e., players). Specifically, this means players should not need to run additional software or perform additional steps to benefit from the system.

#### B. Design Overview

This section describes the design of Servo, a serverless backend architecture for MVEs that provides fine-grained horizontal scalability while maintaining low operational effort from game developers. Figure 4 shows our design, with novel aspects indicated in blue.

Servo is a backend system that modifies and interacts with server instances. It does not interact with the game client, nor change the protocol used between the client and server (addresses **R4**). Instead, Servo modifies existing components in the server to leverage serverless services. To save resources and maintain an update rate of 20 Hz under high workload, Servo uses computational offloading for simulated constructs, a computationally intensive part of the game loop (first part of Requirement **R1**).

*Novel in this work*, Servo leverages serverless technologies such as Function-as-a-Service (FaaS) and managed storage to provide fine-grained scalability for the key MVE components described in Section II-A. We provide first an overview of Servo’s novel elements, and proceed by describing each of these elements in detail in the remainder of this section.

To provide good performance for terrain simulation, Servo modifies the MVE simulator (①) to use a *speculative execution unit* (②) to simulate SCs. The speculative execution unit is responsible for managing speculative executions of each simulated construct, and providing the simulator with the speculative states that turn out to be correct. The execution unit offloads the SCs to serverless functions (③, partially addresses Requirement **R3**). *Unlike the game loop, the function is not required to simulate changes at a fixed rate (e.g., 20 Hz), allowing it to compute state changes speculatively.* Each function invocation simulates a single simulated construct, computes its state changes for multiple simulation steps, and sends these state changes back to the execution unit, allowing the simulator to merge the state changes with those computed in the game loop (partially addresses **R2**).

To independently scale terrain generation, Servo moves the responsibility of terrain generation from the (monolithic) server to an independent serverless function, reducing the server’s computational load (completes **R1**). The *terrain generator* (④) is responsible for invoking terrain generation tasks based on the current location of each avatar. Each generation task (⑤) corresponds to a serverless function invocation, and generates a fixed-sized area of the terrain.

Finally, Servo moves the responsibility of managing persistent game state from the game operator to the cloud provider, storing game state in cloud-based managed storage (⑦, **R3**). MVE persistent storage includes player-, meta-, and terrain-data. We focus in our design on the latter because it is the most data-intensive and is accessed most frequently. To maintain good quality of service (QoS), Servo uses a server-local cache (⑥) with a simple pre-fetching policy to hide latency and performance variability from the managed storage solution (**R2**). A more detailed description of Servo’s serverless storage can be found in our technical report [17, §3.5].

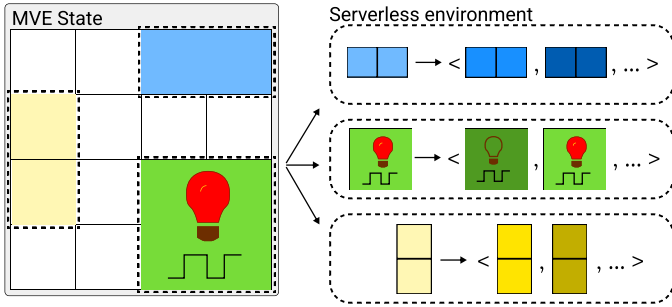


Fig. 5: A visual representation of speculative execution in Servo. Colored areas are simulated constructs. White areas are static parts of the virtual world.

### C. Replicated Speculative Execution for Simulated Constructs

MVEs allow players to build *simulated constructs*. As described in Section II, simulated constructs are part of the virtual world’s terrain, allow players to program the virtual world, and generate additional computational load for the MVE instance. This section describes how Servo uses computational offloading and speculative execution to scale simulated constructs.

Servo scales horizontally by offloading the computation of individual simulated constructs to serverless functions. Whenever a player builds and activates a simulated construct, Servo invokes a function and passes the simulated construct’s current state and the number of steps to simulate. Although this approach scales well with the players’ ability to build an endless number of simulated constructs, it raises an important design challenge: Because simulated constructs are part of the terrain, their update rate must match that of the virtual world (i.e., 20 Hz), and their state changes must become visible to players with low latency (i.e., after a single tick, or 50 ms).

Servo addresses this challenge through replicated, speculative execution. Figure 5 presents an overview of this approach. The left-hand side of the figure depicts a simplified top-down view of a voxel-based 3D virtual world. The colored areas are simulated constructs: simulated elements that change state. An example of a simulated construct is a circuit that powers a lamp, shown in green and red. Servo isolates the state of these elements and forwards each to a serverless function for speculative execution. If no speculative state results are available on the game server on time, or the speculative state is incorrect, Servo falls back to server-side simulation.

Speculative results can be incorrect because, although SCs operate independently of other simulated constructs and players, players have the possibility to interact with the simulated construct, for example by modifying the underlying terrain. Because such changes cannot be part of the original request to the remote function, its results will likely be inconsistent with the new correct state. Addressing this challenge, we include in the request a logical timestamp indicating when a player last modified the simulated construct. The function sends this same timestamp in its reply, allowing the server to disregard the reply if its result is outdated.

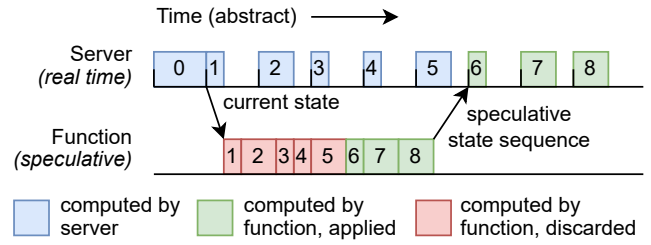


Fig. 6: Example of speculative execution for simulated constructs. Servo simulates the construct locally, until it receives the results from the remote function.

This approach reverses the common application of speculative execution and rollback mechanisms: instead of speculatively applying updates locally and undoing those changes if they turn out to be incorrect, Servo performs speculative execution remotely and only applies changes when they are correct. This approach is feasible due to the game’s fixed update rate, which artificially limits their processing speed.

Figure 6 shows a visual representation of this approach. When a simulated construct is started, Servo starts simulating simultaneously on the server and in a remote function. Starting the simulation on the server hides the (cold-start) latency from the remote function (completes **R2**). However, performing all the work on the server would defeat the purpose of offloading the computation. Instead, we observe that, unlike the server, the function does not need to simulate at a fixed rate, and therefore can (speculatively) work ahead. This allows the function to catch up with, and get ahead of, the game server, and return a collection of future states. Upon receiving a reply from the worker, Servo stops its local simulation and instead applies state updates received from the remote function.

The number of steps that need to be computed locally determines the *efficiency* of the speculative execution. For the example shown in Figure 6, 5 out of 8 steps are computed both locally and in the offloaded function, resulting in an efficiency of  $\frac{8-5}{8} \approx 0.38$ . Although the first speculative execution is likely to reach low efficiency, the efficiency is likely to increase significantly for all following invocations by invoking further speculative execution several steps before the results are needed. Throughout the rest of the article, we refer to this number of steps as the *tick lead*. Returning to the example in Figure 6, the second speculative execution can start at time 6, but simulate starting at state 8, resulting in a tick lead of  $8-6=2$ .

### D. Leveraging FaaS for On-Demand Content Generation

Players exploring the virtual world, or teleporting their avatar to new locations, cause computationally intensive workloads and contention for resources with the game’s simulator, which must maintain stable performance. Here we briefly describe how Servo horizontally scales the MVE’s content generation process by running content generation tasks in remote functions.

TABLE I: Overview of Experiments. Experiments focus on Simulated Constructs (SC), Terrain Generation (TG), and Remote Storage (RS). Components either run locally (L), use serverless computing (S), or combine the two (L+S). For a detailed description of parameters, workload, environment, and explanations on the notation, see Section IV-A.

Focus		Parameters			Workload			Environment		Rep.	Dur. [m]
		SC	TG	RS	Players	Behavior	World	Server	Services		
Section IV-B	SC: System scalability	L+S	L	L	10-200	A	flat	DAS-5	AWS	1	10
Section IV-C	SC: Latency hiding	L+S	L	L	1	-	flat	DAS-5	AWS	1	5
Section IV-D	TG: QoS	-	S	L	5	Sinc	default	DAS-5	AWS	1	5
Section IV-E	TG: System scalability	-	L+S	L+S	100	\$3, \$8, R	default	Azure	Azure	1, 20	10
[17, §4.6]	RS: Perf. variability	-	-	S	8	\$3	default	Azure	Azure	1	10
[17, §4.7]	SC: Performance	S	-	-	1	-	flat	DAS-5	AWS	-	-

TABLE II: Player actions in the random behavior (R). See Section IV-A for a detailed description.

Probability	Action
40%	<b>Move</b> to a random destination at 1 to 8 blocks per second.
30%	<b>Break or place</b> a nearby block.
20%	<b>Stand</b> still.
5%	<b>Send a message</b> to all other players.
5%	<b>Set inventory</b> to a random item.

As described in Section II, MVE’s generate new areas of the terrain on-demand, when players approach these areas. Servo separates the virtual world and content generation responsibilities and moves the content generation component to a remote function. When players approach non-existent areas, the game invokes a remote function for each area that requires generation, providing as parameters the seed for the pseudo-random number generator, and the coordinates of the area. Although the generation of a single area is expected to complete in the order of seconds, all generation requests can be invoked concurrently, leveraging FaaS’s simple scalability.

#### E. Remote State Storage

Whereas MVEs typically save state to local disk, Servo uses serverless storage to reduce the operational effort for the game developer (completes **R3**). However, managed storage can exhibit significant performance variability, which must be hidden from MVE players to provide good QoS. Although Servo uses managed storage for player-, meta-, and terrain-data, we focus in this section on terrain-data, because it is the most challenging: terrain-data consists of relatively large files which are accessed frequently and with strict latency requirements.

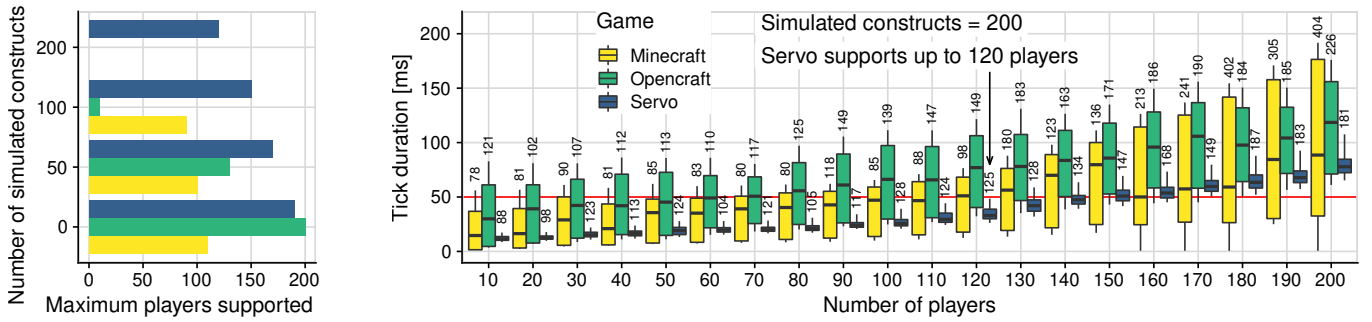
Servo addresses the latency requirements of terrain storage by caching terrain data on local storage and pre-fetching data from remote storage. When avatars change location, the storage service checks which files correspond to the avatar’s current location and the areas in the player’s view range, and load these into memory from the local file system or remote storage. Areas not within any player’s view distance are removed from memory after a certain amount of time. Servo caches these areas on the local file system, reducing the number of accesses to remote storage and saving bandwidth. To improve the cache’s hit rate and hide latency, Servo pre-

fetches terrain data outside of, but close to, the player’s view distance and proactively loads these into memory. In contrast to reads, writes to remote storage are performed periodically.

## IV. EXPERIMENTS

To evaluate our design, we implement Servo on top of Open-craft, an open-source MVE and research platform compatible with the original Minecraft network protocol [12]. We implement Servo’s serverless components on the two most popular commercial cloud platforms: Amazon Web Services (AWS) and Microsoft Azure, and conduct real-world experiments using our prototype implementation of Servo. Table I shows an overview of our experiments, from which we derive the following main findings:

- MF1** Serverless offloading of simulated constructs improves game scalability. Servo supports up to 140 (+1400%) additional players under favorable conditions, compared to state-of-the-art alternatives (Section IV-B).
- MF2** Speculative execution efficiently hides serverless offloading latency, with a median of 84% of speculative states applied without invoking functions in advance, and 100% when invoking functions 10 steps in advance (Section IV-C).
- MF3** Serverless content generation provides good QoS (Section IV-D). Servo can maintain full view distance (128) whereas Open-craft drops below 16.
- MF4** Serverless content generation maintains good performance, increasing the number of supported players by 6 compared to Open-craft (Section IV-E).
- MF5** Caching significantly reduces performance variability when reading game data (see our technical report [17, §4.6]). Servo reduces the 99.9<sup>th</sup> percentile latency for loading terrain data from serverless storage from 226 ms to 34 ms, enabling terrain loading within a single simulation step.
- MF6** Simulated constructs of small and medium scale can be used for speculative execution effectively [17, §4.7]. At least 95% of samples speculatively executing 100 steps for simulated constructs consisting of 252 and 484 blocks simulate at a rate of 488 and 105 updates per second, 24.4 and 5.3 times faster than the simulation rate respectively.



(a) Max. players for increasing SCs. Higher is better.

(b) The tick duration distribution for a varying number of players and 200 SCs. Lower is better.

Fig. 7: Comparing the scalability of Servo with alternative systems. Scalability is expressed in maximum number of players while maintaining stable performance. Missing bars indicate the game could not support any players.

### A. Experimental Setup

Table I shows an overview of our experiments, and details their parameters, workloads, and experiment environment. The experiments cover three major system parameters: *simulated constructs* (SC in Table I, explained in Section II-A and depicted in Figure 2, component 6), *terrain generation* (TG), and *remote storage* (RS). Each of these components either runs locally (L in Table I), uses serverless computing (S), or a combination of the two (L+S).

The experiment workloads are defined through three parameters. First, *players* indicates how many players connect to the game throughout the experiment.

Second, the *behavior* indicates what players do once connected. \$x\$ indicates that players move away from the starting location (i.e., spawn location) at a speed of  $x$  blocks per second, in a star-shaped pattern, whereas \$inc\$ indicates players that increase their velocity over time. These synthetic behaviors increase the amount of terrain covered by the players, and allows stress-testing the game’s terrain generation. R indicates a randomized behavior, in which players perform actions that are common in MVEs. An overview of these actions is available in Table II. We use a randomized behavior because no MVE-specific player behavior models currently exist. A indicates a behavior where players exclusively take move actions, within a bounded area. This behavior limits the amount of terrain generation, and is used to evaluate simulated construct (SC) performance.

Third, the *world* indicates the type of terrain (generation) used. The *default* world type uses procedural content generation and is commonly used by real players and features mountains, rivers, and other natural phenomena. The *flat* world type consists of an infinite plain in all directions. This terrain makes it easy to build, and players commonly use this world type to prototype simulated constructs.

We perform our experiments on the commercial cloud platforms Amazon Web Services (AWS) and Microsoft Azure, and on DAS-5, a local medium-sized supercomputer for research and education [36]. *Server* indicates the location of the MVE server instance, and *services* indicates the platform used for

managed services (e.g., FaaS, managed storage). Unless otherwise indicated, we use 512 MB memory on AWS Lambda, and compare the performance of Servo with Opencraft and Microsoft’s official Minecraft 1.16 Java Edition.

### B. Offloading Simulated Constructs Improves MVE Scalability

Our results show that: (1) Using serverless offloading of simulated constructs, Servo can support up to 120 players under workloads where Opencraft and Minecraft support none (0); and (2) For other workloads tested, Servo can increase the number of supported players under favorable conditions by 40 (+31%) to 140 (+1400%) compared to Opencraft and Minecraft.

Figure 7a shows the maximum number of supported players for varying workloads. The vertical axis shows four workloads with an increasing number of simulated constructs (i.e., increasing computational requirements). The horizontal axis shows the maximum number of supported players, which we define as the maximum number of players for which less than 5% of tick duration samples exceeds 50 ms. The bars indicate different games. Servo (blue) serverlessly offloads the simulation of the simulated constructs. Servo is based on Opencraft (green), which computes simulated constructs locally. Minecraft (yellow) is the official Minecraft server.

The result shows that Servo significantly improves scalability in nearly all cases, and introduces only limited overhead. The workload with 0 simulated constructs shows the baseline performance. Minecraft supports 110 players in this scenario, whereas Servo and Opencraft support 190 and 200 players respectively. For higher numbers of simulated constructs, the maximum number of supported players decreases for all games, but Servo performs best under these conditions.

Opencraft and Minecraft support up to 10 and 90 simultaneous players respectively in worlds with 100 simulated constructs. Servo supports 150 players, an increase of 140 (+1400%) and 60 (+67%) players respectively. When using 200 SCs, Servo supports 120 players whereas Opencraft and Minecraft support zero (0). Even when connecting a single player to Opencraft or Minecraft, their 95<sup>th</sup> percentile tick duration exceeds 50 ms.



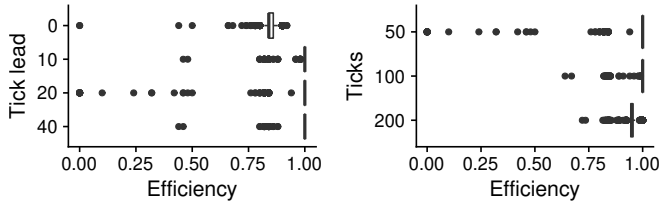


Fig. 8: Efficiency of offloaded simulation for varying tick leads (left), and simulation lengths (right).

Figure 7b shows the detailed performance of all three games when using 200 SCs for multiples of 10 connected players between 0 and 200 (corresponds to the top-row in Figure 7a), and allows an analysis of the observed scalability differences. We further evaluate the scalability of the games for fewer than 10 players, but omit this data from the plot for improved readability. The horizontal axis shows the number of connected players, and the vertical axis shows a boxplot of the game’s tick durations. The whiskers extend to the 5<sup>th</sup> and 95<sup>th</sup> percentiles. The maximum value is shown in the text above each box, and further outliers are omitted. To maintain good and stable performance, the tick duration must be below 50ms. The red line indicates the 50ms threshold.

We observe that for both Minecraft and Opencraft, a significant number of samples exceed 50ms, and that this number increases with the number of connected players. In contrast, Servo’s tick duration is significantly lower, and shows a narrower distribution (box height). Servo supports up to 120 players because fewer than 5% of tick duration samples exceed the 50ms threshold up to this number of players.

We observe that Servo’s tick duration is consistently close to the 25th percentile of Opencraft’s tick duration. Upon further analysis, we observe that both Opencraft’s and Minecraft’s tick duration distributions are bimodal, and that this behavior is caused by their implementation, in which simulated constructs are simulated every other tick. I.e., the bottom half of their distribution shows the performance when not updating the simulated constructs.

Overall, the results show that offloading significantly improves the performance of the MVE.

### C. Speculative Execution Effectively Hides FaaS Latency

Our results show that Servo’s latency hiding mechanism can effectively hide the latency from serverless functions.

Figure 8 shows this result. Both plots show the *efficiency* (Section III-C) of the offloaded computation on the horizontal axis. The efficiency is computed as the fraction of offloaded work that must still be performed by the game server because its result arrives too late at the server instance. The left-hand plot shows the tick lead on the vertical axis, which shows how many ticks in advance of the computation result being required the function is invoked.

The results show that a tick lead of 0 leads to a median efficiency of 84% (top row in the left-hand plot), indicating that 16% of the offloaded computation must still be performed

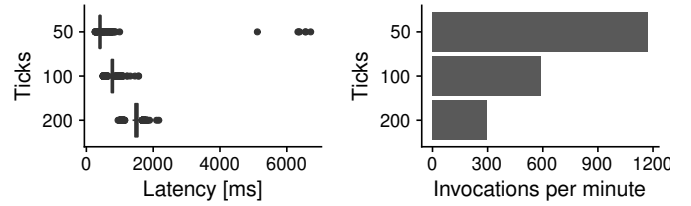


Fig. 9: Latency (left) and number of invocations per minute (right) for varying simulations lengths in ticks.

locally to meet the game’s latency requirements. However, when the function is invoked in advance by 10, 20, or 40 ticks, median efficiency reaches 100% (boxplots align vertically at 1.00). Further analysis shows that for these three configurations, at least 99.1% of invocations have an efficiency of 100%.

The right-hand plot in Figure 8 shows the function efficiency for a varying number of simulation steps, using a fixed 20-tick lead. The plot shows that efficiency remains high (100% median) for both 50 and 100 simulation steps, but drops below 100% for 200 steps. We find the reason for this drop in efficiency is the increased latency of the function when simulating 200 steps, which we explore in Figure 9.

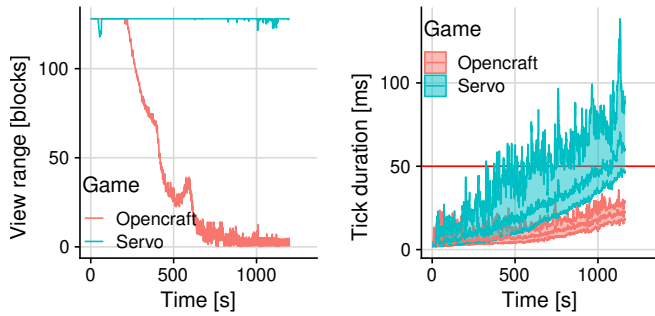
Figure 9 shows the end-to-end latency of invoking a simulation function, and the number of invocations per minute, for varying simulation lengths. The left-hand plot shows that increasing the simulation length increases the function latency. The 200-tick simulation shows a mean latency of 1459ms, which exceeds the lead time of 1000ms (20 ticks of 50ms each). This explains the reduced efficiency in the right-hand plot in Figure 8.

Figures 8 and 9 show numerous outliers in all plots. Further analysis shows that outliers have a strong temporal correlation, providing evidence that these outliers are caused by cold starts. This is surprising because we perform a warm-up iteration of the experiment that uses more invocations per minute than the other configurations, suggesting that AWS starts deallocating function resources within minutes.

Multiplying the mean latency with the number of invocations per minute shows the additional cost of running Servo to be between \$0.216 and \$0.244 per hour. This is comparable to the cost of running one c5n.xlarge instance (\$0.216 per hour), but without the instance management overhead.

### D. Serverless Terrain Generation Provides Good QoS

Procedurally generating terrain for MVEs using serverless functions achieves good QoS. Figure 10 shows this result. The left-hand plot shows how Servo is able to load the required terrain in time, whereas Opencraft does not. The horizontal axis shows time, and the vertical axis shows the minimum distance between a player and the closest missing part of terrain. To provide good QoS, this value should remain at 128, the default configured view distance of the players. The workload consists of five players, each moving in a different direction. The players initially move at a speed of one unit



(a) Distance until closest un- (b) Tick duration over time.  
loaded chunk. Higher is better. Lower is better.

Fig. 10: Serverless Terrain Generation Performance.

per second, and increase their speed by an additional unit per second every 200 seconds.

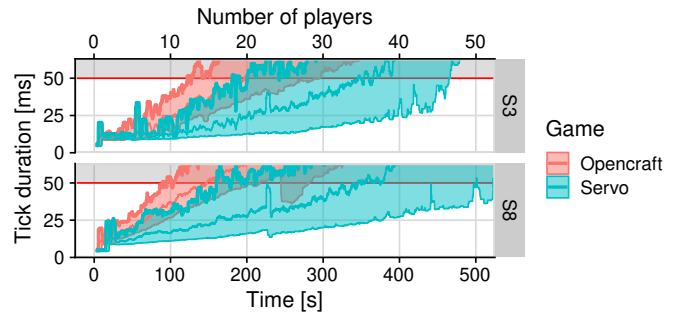
The plots show that Opencraft (red curve) manages to generate sufficient terrain when the players move at 1 unit per second, but fails to do so when their speed increases. After 1000 seconds, when the players move 6 units per second, the closest unloaded terrain is less than 16 units away. I.e., the terrain generation cannot keep up with the workload. In contrast, *Servo maintains good QoS throughout the experiment*. Initially, after roughly 30 seconds, Servo experiences a brief and small performance degradation (dip in the blue curve in the top left). After further analysis, we conclude that this behavior is caused by cold starts. The area around the avatar starting-location is generated locally, before the experiment starts. After 30 seconds, the avatars enter a new area and Servo starts generating terrain serverlessly. After roughly 30 seconds, a sufficient amount of resources has been allocated and Servo’s performance stabilizes.

The right-hand plot shows the tick duration for both games. The vertical axis shows the tick duration, which must remain below 50 ms to meet QoS constraints. The outsides of the colored bands indicate the 5<sup>th</sup> and 95<sup>th</sup> percentiles using a 2.5-second window, and the curve inside the band shows a 2.5-second rolling arithmetic mean. Although Servo maintains good QoS in terms of terrain generation, the 95<sup>th</sup> percentile tick duration starts exceeding the 50 ms threshold while players move at a speed of 2 blocks per second. Further analysis shows that this is caused by the higher number of terrain parts generated for Servo compared to Opencraft. Although content generation happens outside the main game loop for both games, the overhead of loading the content in the game causes overhead which increases tick duration.

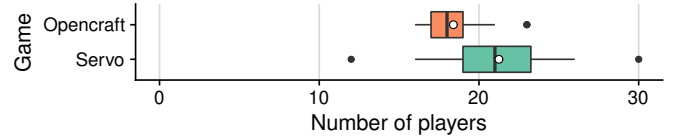
### E. Serverless Terrain Generation Has Good Performance

Under constant workload, serverless terrain generation provides lower tick times than Opencraft, increasing the median number of supported players by 17%, and increasing performance overall. This section presents experimental evidence for this result, for three separate workloads.

Figure 11a shows the tick durations over time for the \$3 and \$8 workloads, in which players move away from the starting



(a) Tick duration over an increasing number of players for workloads S3 and S8. Lower is better. Values must remain below 50 ms to maintain good QoS.



(b) Maximum supported players for the R workload. Higher is better. The white dot indicates the arithmetic mean.

Fig. 11: Performance improvement of serverless terrain generation (Servo) compared to local generation (Opencraft) for varying workloads.

location in a straight line with a speed of 3 or 8 in-game units per second, respectively. Each player goes in a different direction, exploring new areas and generating new terrain.

The horizontal axis on the bottom shows time, and the horizontal axis on the top shows the number of players. As time progresses, more players connect to the game. The vertical axis shows the tick duration, which must remain below 50 ms to meet QoS constraints. The outsides of the colored bands indicate the 5<sup>th</sup> and 95<sup>th</sup> percentiles using a 2.5-second window, and the curve inside the band shows a 2.5-second rolling arithmetic mean.

The figure shows the tick duration increasing for both Opencraft and Servo. This is caused by the increasing workload: every ten seconds, a new player joins the game and starts moving in a new direction. This causes additional terrain to be sent to clients and more terrain to be generated, increasing the load on the server. Under the \$3 workload, Opencraft and Servo support up to 12 and 18 players respectively (the 95<sup>th</sup> percentile curve exceeds the 50 ms threshold). The \$8 workload shows a similar trend, but supports fewer players (9 and 15 respectively) because it is a heavier workload.

Because the plot shows large performance variability (height of colored bands), we perform an experiment with a random behavior workload which we repeat 20 times. The result is shown in Figure 11. The plot shows that Servo supports slightly more players than Opencraft (distribution is more to the right), but has increased performance variability (width of the box is larger). The plot shows that Servo maintains good performance when using FaaS to offload terrain generation.

## V. RELATED WORK

In this section, we present a developing and brief overview of related work. A more detailed discussion of related work is available in our technical report [17]. The benefits of using serverless services have been studied for several novel application domains such as data processing [37], graph processing [38], and video encoding [39]. Furthermore, the application of fine-grained computational units and serverless computing for gaming services has been an active research topic for the past decade [40], [41]. However, *Servo is the first online multiplayer game to improve its virtual-world scalability through serverless computing*. Supporting real-time online-multiplayer games on serverless platforms is challenging because real-time games require consistently high performance and low latency, and because the FaaS programming model (request/reply) does not match well that of real-time online-multiplayer games (message streams).

There exist a variety of architectures for scalable online multiplayer games. An extensive overview is available in [13]. Systems such as VON [42], Colyseus [43], and Rokkatan [44] distribute the simulation of a single static world over multiple machines. In contrast, *Servo* uses additional resources dynamically and is designed specifically to address the computational challenges of MVEs.

Speculative execution techniques have been applied in a variety of domains, such as big data processing [45], state machine replication [46], and in a variety of interactive application types: VNC/SRD [47] uses speculative execution to predict frames for a remote display system. Crom [48] performs speculative execution to hide latency in Web browsing. Outatime [49] uses speculative execution to render frames based on a player's predicted inputs. Although speculative execution is commonly applied close to the user, *Servo* uses speculative execution for computational offloading.

Mirror [50] is an architecture for computation-offloading in mobile games, in which mobile games can offload computation to a remote machine. Whereas Mirror requires the remote machine to be always online, and delays player inputs to hide latency, *Servo* uses serverless functions that only run when the offloaded computation is required and hides latency by performing local active replication until the offloaded simulation is complete.

## VI. CONCLUSION AND FUTURE WORK

Gaming is the world's largest entertainment industry, and MVEs are the best-selling game of all time. Minecraft, the canonical example of an MVE, has more than 130 million active monthly players. However, due to the inherent complexity of *modifiable* virtual worlds, this scale can only be reached by replicating large numbers of isolated instances, preventing players from playing together. In this work, we explore the use of serverless computing to address the complexity and limited scalability of MVE instances. Specifically, we design *Servo*, a prototype MVE which leverages serverless to offload both compute and storage from MVE server instances to the cloud. Our results show that *Servo's* offloading increases the number

of supported players by 140, and that it can successfully hide (cold-start) latency from players.

In future work, we aim to investigate how to scale large simulated constructs while meeting their QoS requirements, and (serverless) MVE designs without a central server.

## ACKNOWLEDGMENTS

This work is funded by the projects NWO OffSense, EU Graph-Massivizer, CloudStars, and by structural funds from VU Amsterdam. We thank Tim Hegeman, who helped us improve this work by providing valuable technical expertise.

## REFERENCES

- [1] D. Bar-El and K. E. Ringland, "Crafting game-based learning: An analysis of lessons for minecraft education edition," in *FDG*, 2020.
- [2] Michele Melchiorre, Senior VP BMW, "The BMW iFactory - Highly Efficient and Competitive," in *Keynote ISC*, 2022.
- [3] "Minecraft officially crosses over 141 million monthly active users," <https://www.windowscentral.com/minecraft-live-2021-numbers-update>, October 2021.
- [4] van der Sar *et al.*, "Yardstick: A benchmark for minecraft-like services," in *ICPE*, 2019.
- [5] Mojang, "Minecraft Realms," Jan. 2022, "You and up to 10 friends can play together at one time." [Online]. Available: <https://www.minecraft.net/en-us/realms>
- [6] B. Dear, Ed., *The Friendly Orange Glow: The Untold Story of the PLATO System and the Dawn of Cyberculture*. Pantheon Books, 2017.
- [7] Newzoo, "Newzoo global games market report 2021 free version newzoo," July 2021.
- [8] Donkervliet *et al.*, "Towards supporting millions of users in modifiable virtual environments by redesigning minecraft-like games as serverless systems," in *HotCloud*, 2020.
- [9] Eyk *et al.*, "Serverless is more: From PaaS to present cloud computing," *IC*, vol. 22, 2018.
- [10] J. Schleier-Smith *et al.*, "What serverless computing is and should become: the next phase of cloud computing," *Commun. ACM*, vol. 64, no. 5, 2021.
- [11] P. C. Castro *et al.*, "The rise of serverless computing," *Commun. ACM*, vol. 62, no. 12, 2019.
- [12] Donkervliet *et al.*, "Dyconits: Scaling minecraft-like services through dynamically managed inconsistency," in *ICDCS*, 2021.
- [13] González *et al.*, "Key technologies for networked virtual environments: A new taxonomy," *CoRR*, vol. abs/2102.09847, 2021.
- [14] Liu *et al.*, "Survey of state melding in virtual worlds," *ACM Comput. Surv.*, vol. 44, 2012.
- [15] Liu and Theodoropoulos, "Interest management for distributed virtual environments: A survey," *ACM Comput. Surv.*, vol. 46, 2014.
- [16] Gilmore and Engelbrecht, "A Survey of State Persistency in Peer-to-Peer Massively Multiplayer Online Games," *TPDS*, vol. 23, 2012.
- [17] Donkervliet *et al.*, "Servo: Increasing the scalability of modifiable virtual environments using serverless computing – extended technical report," *arXiv*, 2023.
- [18] Worsley *et al.*, "Multicraft: A multimodal interface for supporting and studying learning in minecraft," in *HCI-Games*, vol. 12790, 2021.
- [19] Slovák *et al.*, "Mediating conflicts in minecraft: Empowering learning in online multiplayer games," in *CHI*, 2018.
- [20] Zhu and Heun, "Teaching and learning of chinese history in minecraft: A pilot case-study in hong kong secondary schools," in *IDC*, 2017.
- [21] "Minecraft Official Site — Minecraft Education Edition," <https://education.minecraft.net/en-us/homepage>.
- [22] A. Natividad, "How Greenpeace Used Minecraft to Stop Illegal Logging in Europe's Last Lowland Primeval Forest," *Adweek*, Jan. 2018. [Online]. Available: <https://bit.ly/MinecraftGreenpeace>
- [23] "The Uncensored Library – Reporters without borders." [Online]. Available: <https://www.uncensoredlibrary.com/>
- [24] "Steam Search," Jan 2021. [Online]. Available: <http://bit.ly/steam-mves>
- [25] "Mods - Minecraft - CurseForge," Jan 2021. [Online]. Available: <http://bit.ly/ModsForMinecraft>
- [26] J. Mann, "Meta's spent \$36 billion on the metaverse. The iPhone and Xbox cost way less." October 2022.

- [27] N. Stephenson, Ed., *Snow Crash*. Bantam Spectra Books, 1992.
- [28] Eyk *et al.*, “Beyond microbenchmarks: The SPEC-RG vision for a comprehensive serverless benchmark,” in *ICPE Companion*, 2020.
- [29] Yu *et al.*, “Characterizing serverless platforms with serverlessbench,” in *SoCC*, 2020.
- [30] K. Raaen and A. Petlund, “How much delay is there really in current games?” in *MMSys*, 2015.
- [31] Z. Ivkovic *et al.*, “Quantifying and mitigating the negative effects of local latencies on aiming in 3d shooter games,” in *CHI*, 2015.
- [32] Nae *et al.*, “Dynamic Resource Provisioning in Massively Multiplayer Online Games,” *TPDS*, vol. 22, 2011.
- [33] S. Kounev *et al.*, “Toward a Definition for Serverless Computing,” in *Serverless Computing (Dagstuhl Seminar 21201)*, 2021, vol. 11.
- [34] V. Nae *et al.*, “Efficient management of data center resources for massively multiplayer online games,” in *SC*, 2008.
- [35] M. Claypool and K. T. Claypool, “Latency and player actions in online games,” *Commun. ACM*, vol. 49, no. 11, 2006.
- [36] Bal *et al.*, “A medium-scale distributed system for computer science research: Infrastructure for the long term,” *Computer*, vol. 49, 2016.
- [37] Jonas *et al.*, “Occupy the cloud: Distributed computing for the 99%,” in *SoCC*, 2017.
- [38] Toader *et al.*, “Graphless: Toward Serverless Graph Processing,” in *ISPDC*, 2019.
- [39] Fouladi *et al.*, “Encoding, fast and slow: Low-latency video processing using thousands of tiny threads,” in *NSDI*, 2017.
- [40] Bernstein *et al.*, “Orleans: Distributed virtual actors for programmability and scalability,” Tech. Rep., 2014.
- [41] B. Bors, *Game Backend Development*. Apress, 2023.
- [42] Hu *et al.*, “VON: A scalable peer-to-peer network for virtual environments,” *IEEE Netw.*, vol. 20, 2006.
- [43] Bharambe *et al.*, “Colyseus: A distributed architecture for online multiplayer games,” in *NSDI*, 2006.
- [44] Müller *et al.*, “Rokkatan: Scaling an RTS game design to the massively multiplayer realm,” in *ACE*, 2005.
- [45] Xu *et al.*, “Chronos: A unifying optimization framework for speculative execution of deadline-critical MapReduce jobs,” in *ICDCS*, 2018.
- [46] Wester *et al.*, “Tolerating latency in replicated state machines through client speculation,” in *NSDI*, 2009.
- [47] Lange *et al.*, “Experiences with client-based speculative remote display,” in *ATC*, 2008.
- [48] Mickens *et al.*, “Crom: Faster web browsing using speculative execution,” in *NSDI*, 2010.
- [49] Lee *et al.*, “Outatime: Using speculation to enable low-latency continuous interaction for mobile cloud gaming,” in *MobiSys*, 2015.
- [50] Jiang *et al.*, “A mirroring architecture for sophisticated mobile games using computation-offloading,” *Concurr. Comput. Pract. Exp.*, vol. 30, 2018.