# MATCHMAKING SERIES: PING

*CALL OF DUTY – APRIL 4, 2024*

## Overview

The design of *Call of Duty* matchmaking is driven by our goal to create the best experiences for our player community. We believe that getting matchmaking right for our players means they enjoy the multiplayer experience more, play longer, quit less, and have more fun.

On January 28, 2024, the *Call of Duty* blog [1] published a general introduction to the *Call of Duty* matchmaking process detailing several factors used to build online matches. For reference, this is the detail provided in the franchise blog:


1. CONNECTION – As the community will attest, Ping is King. Connection is the most critical and heavily weighted factor in the matchmaking process.
2. TIME TO MATCH – This factor is the second most critical to the matchmaking process. We all want to spend time playing the game rather than waiting for matches to start.
3. The following factors are also critical to the matchmaking process:
    - PLAYLIST DIVERSITY – The number of playlists available for players to choose from.
    - RECENT MAPS/MODES – Considering maps you have recently played on as well as your mode preferences, editable in Quick Play settings.

- SKILL/PERFORMANCE– This is used to give our players – a global community with a wide skill range – the opportunity to have an impact in every match.
- INPUT DEVICE – Controller or mouse and keyboard.
- PLATFORM – The device (PC, Console) that you are playing on.
- VOICE CHAT – Enabled or disabled.

The publication of the *Matchmaking Series: Ping* white paper launches the first in a series our team is drafting to expand information about *Call of Duty's* matchmaking process. Our next entry in this series will examine how skill plays a role in the process.

For the purposes of this white paper, our focus will be on the core components of the most heavily weighted factors discussed in the franchise blog: Ping as it relates to CONNECTION and Search Time as it relates to the process of TIME TO MATCH.

- *Ping:* The round-trip time for data to travel from the player's game client to the dedicated server running the game simulation. Lower ping means a smoother in match experience, as there is less delay in the game simulation.
- *Search Time*: The time in seconds between when a player starts a search, and when we provide them with a lobby to connect to. The system is designed to get players into a match quickly; we do not want to leave players waiting too long.

Ultimately any matchmaking factors trade-off against one another, and we are constantly tuning our matchmaking algorithm to find the right balance. For example, using the above factors: if we spend longer searching for a match, we're more likely to find one with a low ping and Skill Disparity, but spending too long searching is a negative player experience, and doesn't guarantee a better result.

We must also take into account varying player populations across the world. It takes longer to form a game in a low-population region like Antarctica vs a high-population region like New York.

We also have to adjust our tuning based on factors external to matchmaking, such as game design, and evolutions in the broader internet ecosystem. Though this is not the focus of this document.

To be sure we're providing the players with the best in-match experience, we pay close attention to a blend of matchmaking-focused key performance indicators (KPIs), such as Delta Ping, lobby and match fullness, skill disparity, match outcome metrics and search time, as well as more player-focused KPIs, including hours-per-user (HPU), lobby quit rates, player retention/churn and player survey results. These metrics are not used directly as inputs to the matchmaking selection algorithm but are observable outcomes that tell us if our matchmaking approach is working in the eyes of our players.

This document will focus on the most important factor, ping. We will talk about why it is important, how data center selection works, how ping is used in matchmaking and how we measure our success. It is intended for professionals building and deploying low-latency multiplayer titles, and we hope it spurs greater discussion and sharing of approach among peers in the industry.

## Terminology

*Dedicated Server* - A game server running in a data center.

*Delta Ping* - The difference in packet round trip time between a player's best data center and the data center they are playing on

*Party* - A group of one or more players, playing together as an atomic group.

*Lobby* - A collection of parties, hosted by a dedicated server, that are either about to play a match, in the process of playing a match, or in the process of finishing a match

*Core MP* - The classic *Call of Duty* multiplayer games modes such as *Team Deathmatch* and *Domination* with 10 to 12 players, as opposed to *Call of Duty: Warzone*, *DMZ*, *Zombies*, etc.

*Playlist* - A collection of game modes and maps.

*Tick rate* - Frequency at which a game server updates the game state

*DC* - Data Center

*BR* - Battle Royale, the original mode included in *Call of Duty: Warzone*

## Matchmaking Ping

*Call of Duty* uses a client-server model for hosting matches, where the client is the player's console or PC, and the server is running in one of many data centers around the world. The latency from client to game server is a significant factor in the quality of the player experience, so the *Call of Duty* matchmaking system aims to minimize ping as much as possible, while still finding a good match in a reasonable amount of time.

There are two different values of "ping" when talking about game servers:

1. The network latency from client to server
2. The in-game latency, that is the time from client update to server response while in game

The in-game latency depends on the simulation implementation and tick rate in the game server, as well as the frame rate of the client and so is beyond the control of the matchmaking

system. The matchmaking system aims to minimize the network latency and is the primary topic of discussion below.

## Impact of Ping on Players

As the server runs the game simulation, player actions are not “complete” until they have been sent to the server, processed, and the client has been updated with the new game state. The total duration of this process consists of the network latency, and the simulation tick time in the server and the client. Higher network latency will increase the delay for the client.

First person action games like *Call of Duty* are particularly sensitive to latency. In a turn-based game a latency of 200 milliseconds would not be a problem, but in *Call of Duty* this would be considered a poor experience, and the player could be at a disadvantage compared to players with lower pings.

The *Call of Duty* netcode [2] aims to reduce the perceived impact of latency, but it can’t eliminate it. High pings can cause anomalies such as other players warping or teleporting, or apparently accurate shots missing the target.

This can cause a frustrating player experience. Some players may be sensitive to even small increases in latency.

## Searching Population and Matchmaking Wait Time

Any matchmaking system will take some amount of time to find a match. Fundamentally, this is tied to the number of players searching at the same time, who have compatible search criteria.

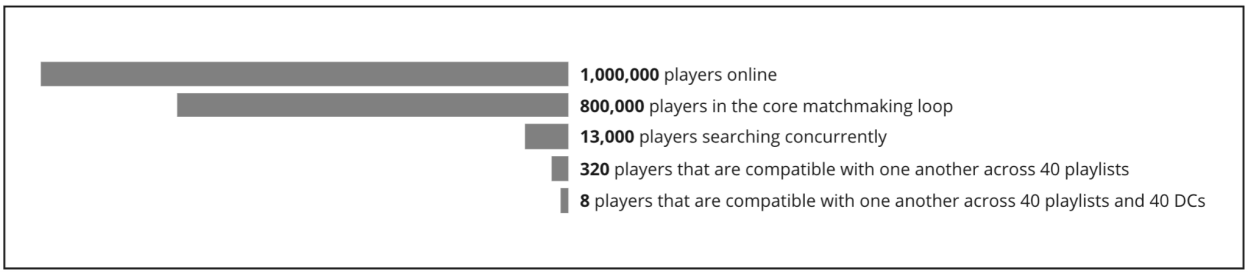
Matchmaking wait time (or search time) is not the full time window between the button press of the start search button and landing in a match, ready to play. Matchmaking search time is limited to the time between the start search button and computing a resulting lobby to place a player into. This doesn’t include the orchestration process of connections, the level load time, or any other overhead associated with actually getting a player into a match. These are all important things to consider about the user experience, but our focus here is on the Matchmaking wait time.

A system with few constraints on matching, and a high population may find matches within seconds, and a matchmaking system with strict constraints may take minutes.

Even a singular focus on ping causes search times that are non-trivial.

Let’s consider an example of 1 million online users in a game. In most games, players do not spend all their time in the core loop of matchmaking, playing a match, then matchmaking again. Let’s say for our example, players spend about 80% of their time in this core loop. So our 1 million players are actually 800,000 in the core loop. Suppose a match takes 15 minutes to play.

Ideally, the matchmaking time would be a short 15 seconds, as we want this to be far smaller than our time spent in a match. This means only 1 in 60 players are searching at any given moment, so 800,000 online players in the core loop are only 13,000 searching concurrently. These 13,000 searches are split (in the case of *Call of Duty*) across approximately 40 Data Centers (DC), and about 40 Playlist options in our Core MP offering. This leaves on average about 8 players that are compatible with one another during any given 15 second interval. Keep in mind the average match size in *Call of Duty* Core MP is 12. So, there isn't even a full match. This average of 8 is a drastic oversimplification, but it illustrates how even with large numbers of concurrent players, ideal ping isn't always practical if we want to match players quickly.




*Not-to-scale depiction of the number of compatible players during any given 15 second interval. Used for example purposes only.*

Players are not evenly distributed across all of our DCs and Playlist offerings. *Team Deathmatch* is very popular and so will have a faster matchmaking time. Other modes, such as *Hardcore Search and Destroy*, will have fewer interested players resulting in longer wait times and higher ping as a result. A similar idea also applies to DCs. Depending on the time of day, one part of the world may have a very high player count while the other side of the world will have a very low player count. So, if you're looking for a match at 3am local time, expect to spend more time waiting for a match and higher ping times, to find compatible players.

In games with strict rules on rank matching, matchmaking can take much longer. This is, again, driven by the population of available players. Taking the above example, even if a game attempted to match players within 25% of the player population in terms of rank, those million online players quickly turn into 2 compatible players every 15 seconds, on average. Add in the constraint of roles which may be played in a match [3], and the problem gets even more complex. Ultimately, this increases the wait time even further, or requires other tradeoffs.

We try to make matchmaking times as short as possible, and in practice this means the process usually averages around 30 seconds. Depending on the region, playlists and times of day, this may be much faster or slower. Ranked modes also have a longer search time, as their matchmaking is also constrained on player divisions.

Wait time is an important consideration when optimizing ping, as excessively long wait times will lead to players canceling searches or quitting the game. Matchmaking in *Call of Duty* is tuned to handle varying search volumes for different times of day and different regions dynamically. This system manages the tradeoffs between search time and player ping.

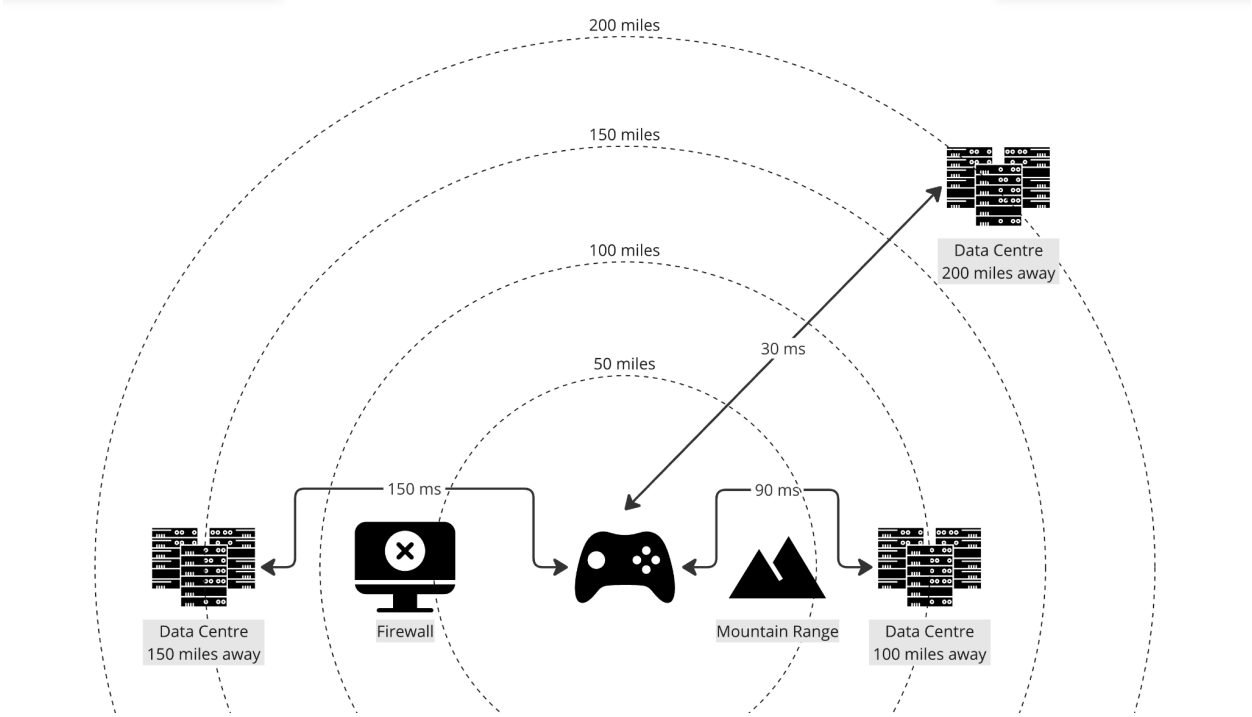
## Identifying Good Data Centers for Players

Before considering how to match groups of players, we must understand what we are trying to achieve for an individual player.

With a large number of data centers around the world, we could potentially place a given player into a match running in any of them. If we pick a data center on the other side of the world, then they will likely have a very high ping. Obviously, we would ideally pick the data center to which the player would have the lowest ping.

To do this we need some idea of the player's pings to the data centers. *Call of Duty* has, in the past, used IP-based geolocation combined with physical distance to give an approximation of expected ping.

Physical distance to the data center doesn't always correlate with ping, however. This is because the internet connections between two physical locations are generally not in straight lines, and many factors other than distance affect the achieved latency. For example, physical features such as mountains or seas may prevent direct cable connections. In other cases, commercial considerations, and various other circumstances prevent connections between neighboring countries, force longer routes than necessary, or add delays due to firewalls etc [4].




The data center with the best ping may not be closely related to distance

The *Call of Duty* game client runs quality-of-service checks against all data centers when the game is started, and periodically between matches. This allows us to use real ping and packet loss data for each data center from the player's perspective.

Given these ping results, we can identify the data center with the lowest ping for a player. We don't only consider matching players into their best DC. Other data centers may have very similar pings to the player's best data center. Often, players have several DCs that are equally good from a player experience perspective. In cases where the search population is low, we may ultimately trade off ping in favor of finding a match quickly.

For example, a player with a ping of 16 milliseconds to a data center in Germany might have a ping of 18 ms to Holland, and ping of 22 ms to France. The first two will be effectively the same, and the next will be very similar, and a good option if using it reduces wait time or improves some other matchmaking criterion.

Based on this idea, the *Call of Duty* matchmaking process aims to reduce the difference between the best data center ping, and the achieved data center ping. We call this difference “Delta Ping”.

## Data Center Locations

A player’s best possible ping is limited by the available data centers. The more data centers available in different locations, the more likely that a player will have a good ping to at least one. But a data center in an area with a low search population might not have enough players to actually form games. As an extreme example, a data center in Antarctica would give good pings for the scientists living there, but it is unlikely that they could form a *Call of Duty: Warzone* match.

## Matching Players with Data Centers

The matchmaking system creates matches consisting of sets of players, which are assigned to servers in a DC. Given the goal of minimizing the difference to each player’s best data center ping, while still satisfying other matchmaking constraints, how should we group players?

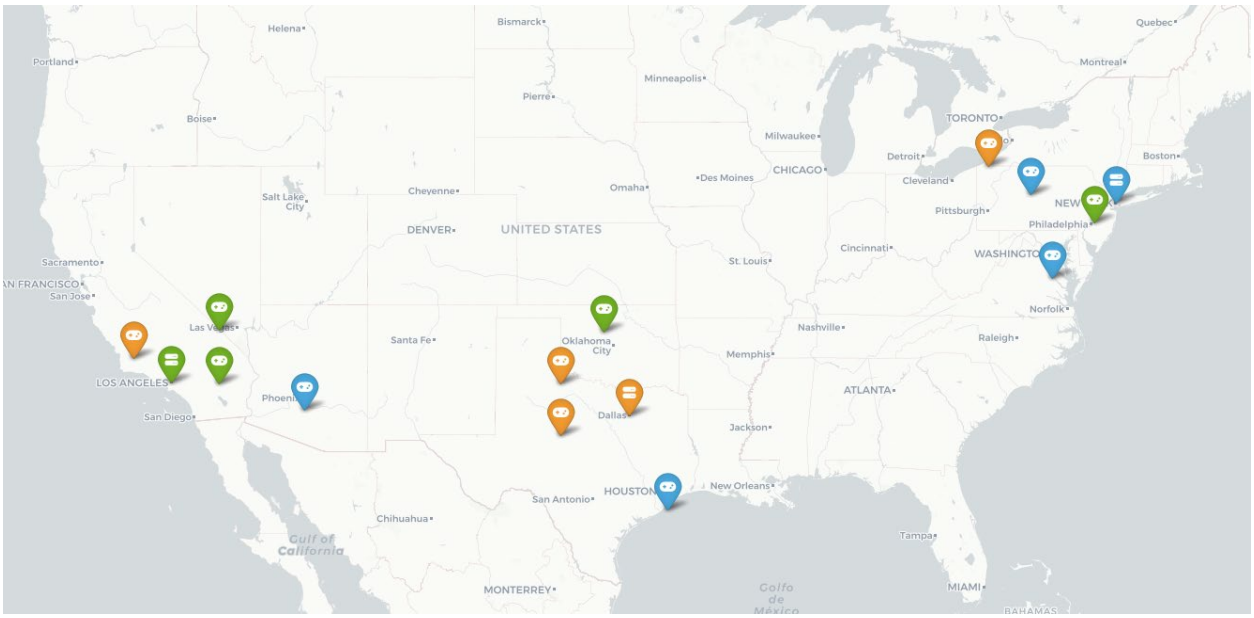
One option would be to assign players solely to their best data center and have a separate queue per data center. This would work reasonably well for busy data centers and popular and low player count modes. But if you searched for a ranked match at 4 in the morning then you could have an excessively long wait time, or never form a match at all, when accepting 5 milliseconds higher ping could allow a larger pool of players to match. The more data centers you have, the worse this option gets, as you split the searching population even more.

Another option is to select a set of players based on other matchmaking criteria (such as region, language, rank, skill etc.), and then try to assign that match to a data center. If we select the players solely based on criteria such as rank, then it is quite likely that they will not share a common data center with playable ping.

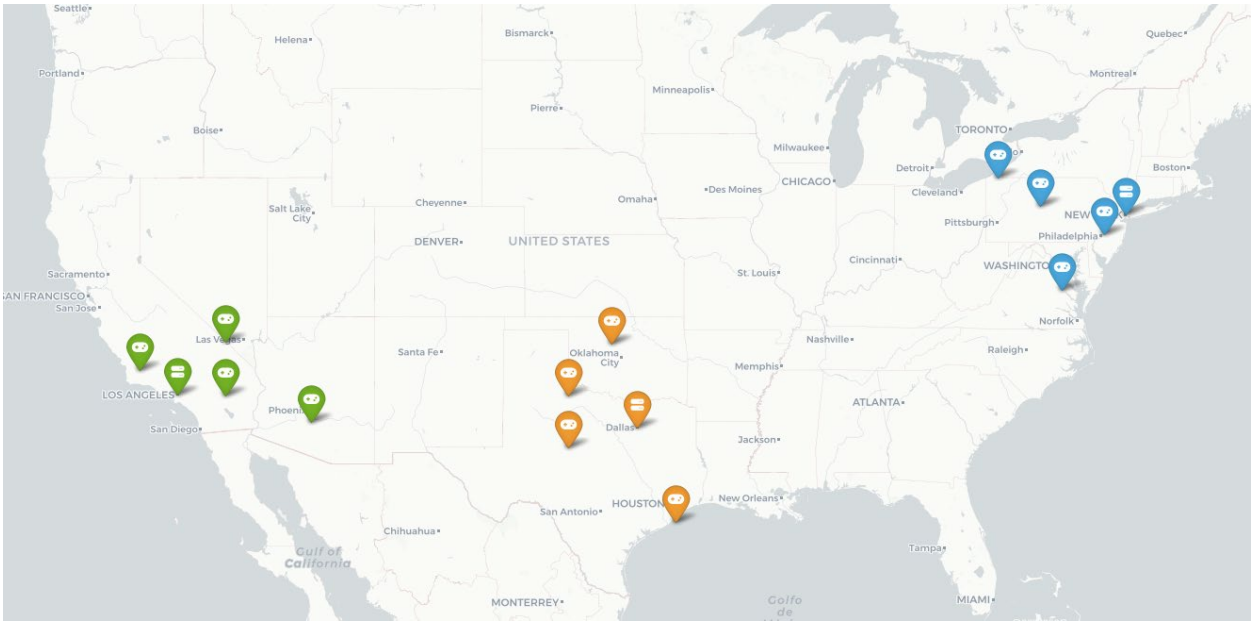
A game could aim to improve on this by using Matchmaking *regions*. Players either manually select a region or are automatically placed in a region by using IP geolocation or ping thresholds. The matchmaking system groups players without consideration of ping. Then the formed matches are assigned to data centers based on something like the minimum average ping of players in the match to the data center.

This approach has two main downsides: poor optimization of ping within regions, and suboptimal matching for players close to border regions.

For example, say we have a North America matchmaking region, and players searching in Los Angeles and New York. All of these players are in the same region, and so can match with each other. So, the formed matches are likely to have players from both cities. To minimize the average ping, it chooses a server running in Dallas. This will typically add around 30 ms of ping, compared to the best data center. Data centers on the East and West coasts are poorly utilized. If a player is matched into a data center on the opposite coast, then it will typically add around 60 ms ping.

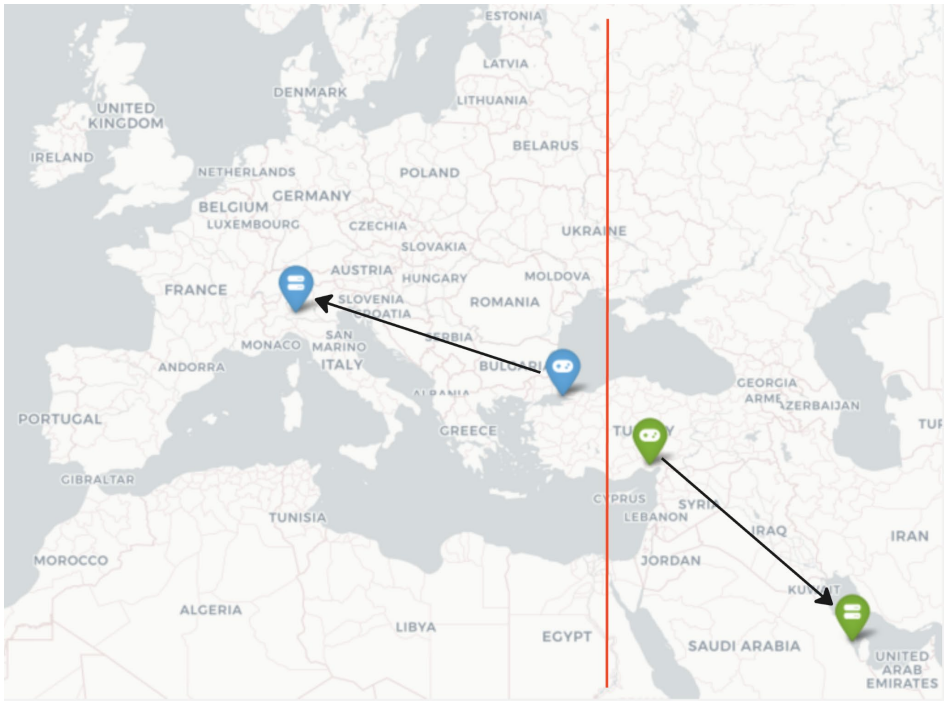


Naive Regional Match formation




Match formation with grouping by data center ping

For the border issue, consider two hypothetical regions: Europe and Asia. Players in Turkey could potentially use data centers in Europe or in the Middle East. By having regions, we restrict them to one region, so they have less population to match with, making other matchmaking KPIs worse. At different times of day there might be more searching population in one region than another, possibly forcing players to manually change regions.




Region selection splits the search population and reduces match quality

The first issue, where we end up just using the data center in the middle of the region, can be alleviated by splitting the regions: e.g. into East and West NA. But this makes the border issues worse. Now players in the Midwest are forced into one region or the other.

*Call of Duty* uses a different approach to solve this problem, which we will describe in the next section.

## Acceptable Data Center Backoff

The *Call of Duty* matchmaking system collects a large number of searching players and attempts to group these players into matches that satisfy the matching constraints.

Some of these constraints are fixed: for example, a player searching for the *Resurgence* game mode may not match with a player searching for *Team Deathmatch*. Other constraints are relaxed, or backed off, over time spent searching.

Data center ping has a fixed component and a backoff. The fixed component is an absolute cutoff beyond which the search will not match. This cutoff can vary depending on the location and search population. For example, players in North America with good pings have a much lower cutoff than players in Africa due to the available search population and data centers.

The backoff based constraint uses increments of Delta Ping over search time. The backoff steps used may vary depending on game mode, location, and search population. The matchmaking system does this to manage the differing population demands of different modes and regions. To form a match, all players must have a common acceptable data center based on their current backoff level. This time-based backoff generally means that when there is a high search population, then players will match quickly, and will get their best data center, or one very close to it. If the search is waiting for a long time, the set of acceptable data centers will expand, and the player may be placed in a data center with higher ping.

After a match has been formed, it must be assigned to a data center. The data center to use is selected from the set of currently acceptable data centers, based on each player's current backoff levels. Within the set of acceptable data centers, we pick the one with the lowest average ping.

## Matching Algorithm

In the section above we discuss the rules for matching groups of players, but not how these rules are used. In this section we will give an overview of the *Call of Duty* matching algorithm and heuristics.

A *search* in the matching algorithm is either a party searching for a lobby, or an existing lobby searching for additional players to fill free spaces. A lobby searching for additional players is

treated very similarly to a party searching for a lobby, but it has only one acceptable data center, as it will already have been assigned a dedicated server.

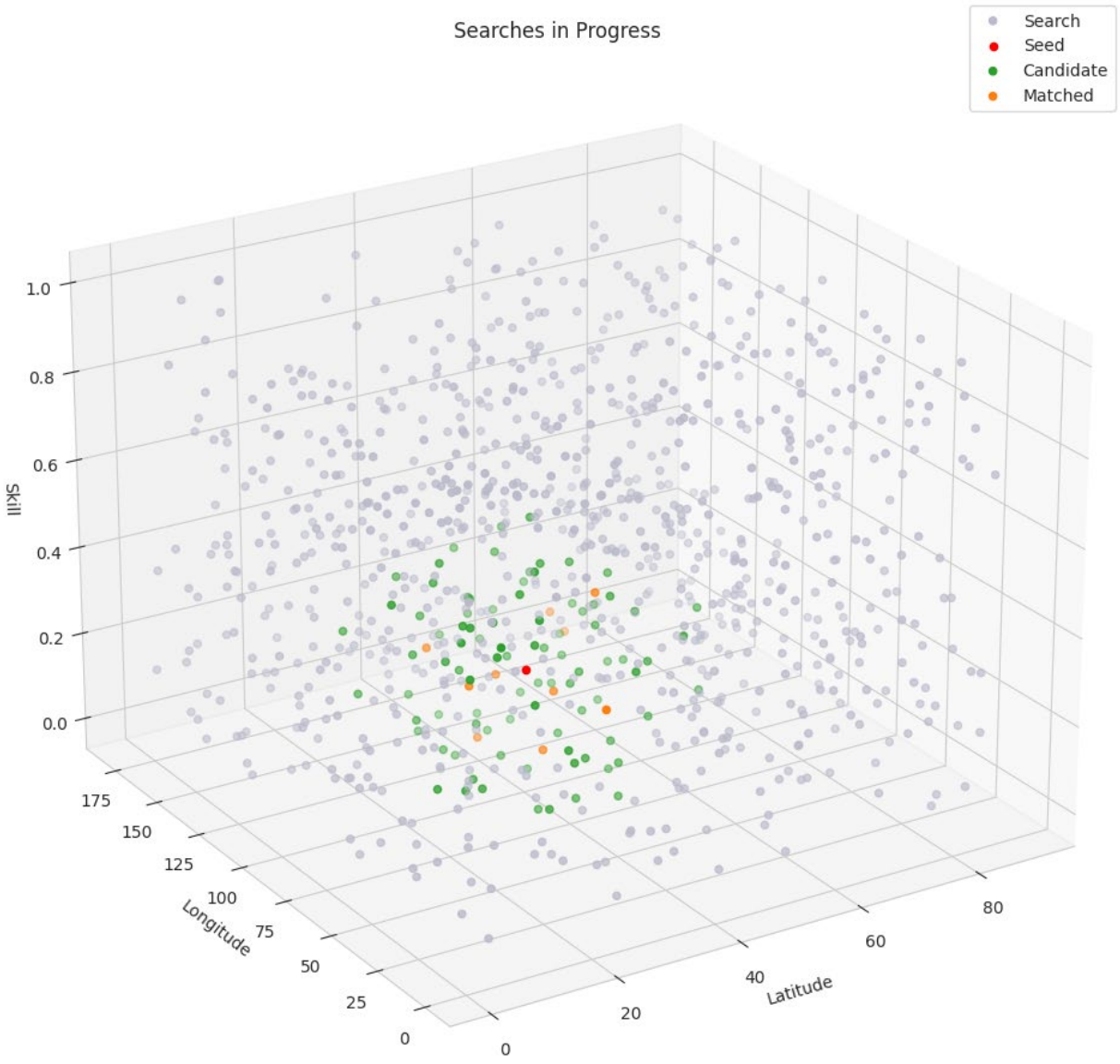
The matching algorithm runs on a fixed interval, typically every 5 seconds, on all searches currently in progress. This progresses roughly in a series of steps:

1. Some searches are used as *seeds*. For some game modes all searches can be seeds, and for others a random subset of searches is used.
2. For each seed an ordered list of *candidate* searches is found using a heuristic. This heuristic looks at a large number of matching factors including physical distance, platform, control scheme, skill etc. This keeps the computational complexity far lower than considering the entire searching population at any given time and has been found to have little to no impact on match quality.
3. A greedy algorithm is used to add candidate searches to a possible match for the seed, such that the matching constraints are satisfied, and a quality score is maximized for each candidate added. This is where the bulk of the matchmaking conditions are considered
4. If a given seed finds a set of candidates that satisfy the constraints for all searches in the set, then a match is created, and the seed and candidates are removed from the pool of searches.

The throughput of the matching process is inversely proportional to the number of candidates considered for each seed, so finding a good set of candidates is important.

For example, we could have:

- 4000 searches available for matching
- 1000 of those selected for use as seeds
- 400 candidates per seed
- Producing matches with 100 players




Simplified view of the search space

As previously discussed, geographical distance is not a perfect approximation for data center pings, so the use of distance in the candidate heuristic may not provide optimal candidates for data center matching. In practice, with a large enough candidate set, it has been a good enough approximation.

The heuristic can also consider set overlap for configurable fields. *Call of Duty* provides a “Quick Play” option where players can select multiple modes that they would like to play. Maximizing the overlap between the selected sets of modes for different searches improves the matching algorithm’s ability to find a common acceptable mode.

## Server Capacity

Data centers may have a variety of hardware, with different specs and configuration, to accommodate the needs of the various game modes. Larger data centers tend to be located in high population areas where there is high demand for game servers. If the selected best data center does not have an available idle server for a match, then we will attempt to use the next best data center, and so on. If there is no available server in a common acceptable data center, then the player searches will be restarted.

A separate system manages the number of servers running in each center to keep enough idle servers available for new matches. The matchmaking system is periodically updated with the available capacity per data center, so that it can avoid creating matches in a data center with low or no capacity. If none of a player’s acceptable data centers have any idle servers then a backoff request will be returned to the game client, and it will retry after a delay. This is to protect the matchmaking system during major outages. This backoff may also be applied during a rollout of a new server and client release, where the game client does not match the currently deployed servers.

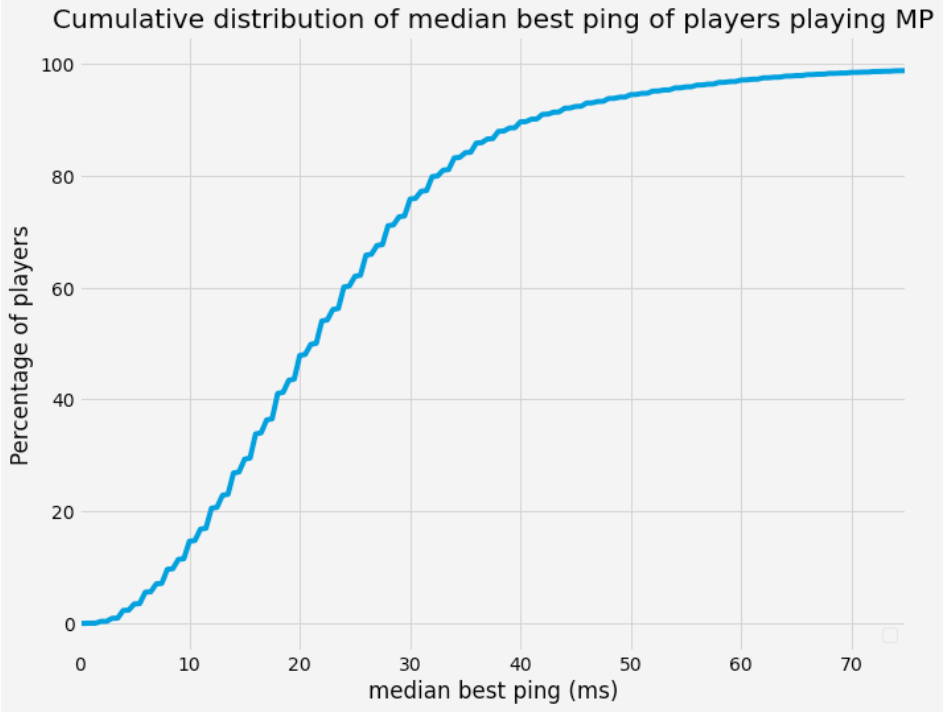
## Monitoring and Results

The ping achieved by *Call of Duty* matchmaking is monitored to ensure that players are getting good results. The location and use of data centers is also monitored to ensure that they are being effectively used. Whenever matchmaking criteria are added or changed, or game modes are changed, the impact on players’ pings is analyzed.

The primary focus of our attention is on Delta Ping, which is discussed below. We also monitor in-game latency, but as much of that time is outside the control of the matchmaking system it isn’t the focus of our attention.

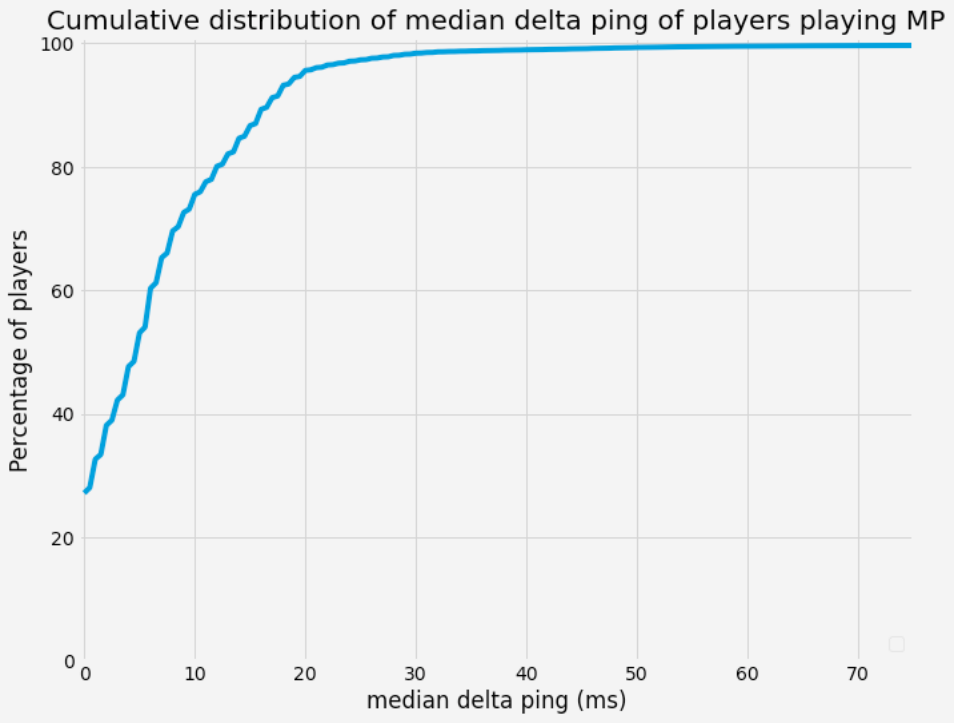
Ping measurement is handled by the game client. Each game client will take several ping measurements periodically to all of our DCs worldwide, between games. This ensures the matchmaking system has up to date measurements for all players at all times. The matchmaking system then records these measurements as part of each search, which in turn feeds into our monitoring of each and every search in *Call of Duty*.

Thanks to the large number of data centers used around the world, 94% of *Call of Duty* players have a best data center ping of 50 milliseconds or lower.




Cumulative distribution of player best data center pings

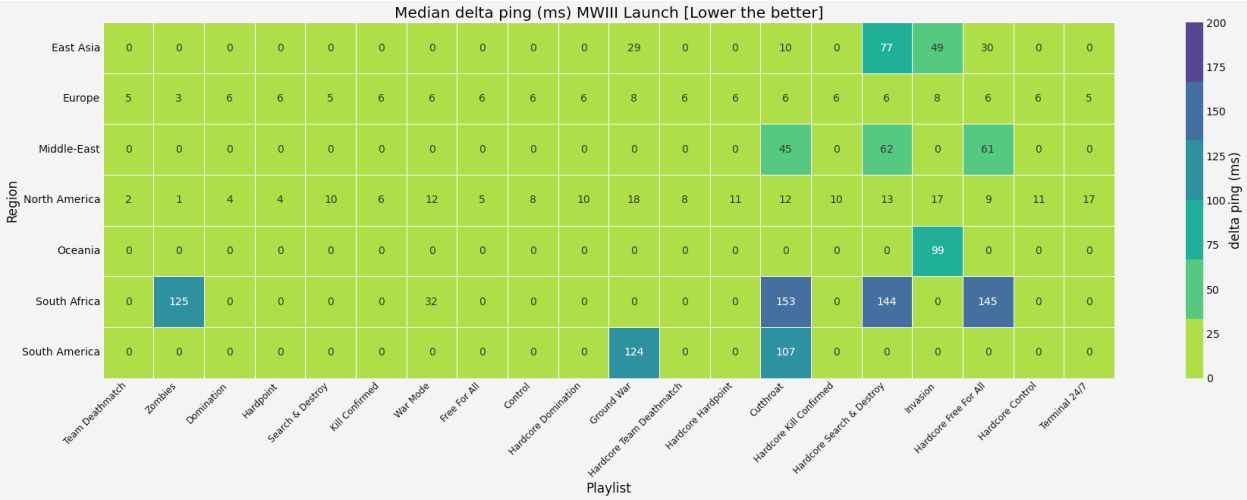
The matchmaking system described above is effective in achieving good pings for players. 75% of players get results which are within 10 milliseconds of the best possible data center. 95% are within 20 milliseconds of the best possible data center.




Cumulative distribution of additional ping from best data center in results

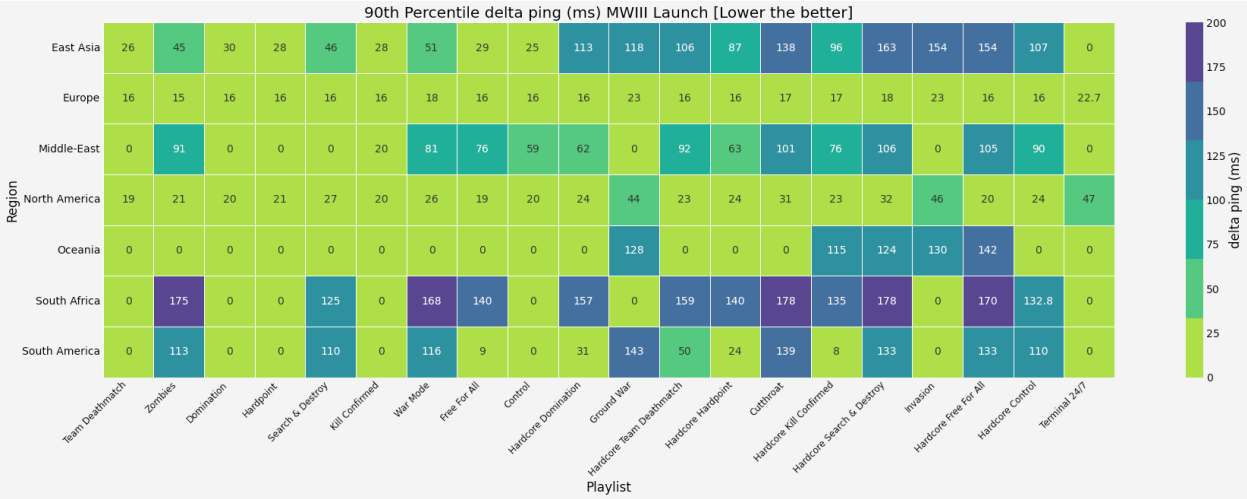
The higher ping values in this distribution are generally from parts of the world with low search volumes, or are from off-peak times of day.

Activision actively monitors Delta Ping regionally and across all playlists. Taking a snapshot of Nov 9th to 15th 2023 (just after the launch of *Call of Duty: Modern Warfare III*), we can see a breakdown of player Delta Ping below:




The median shows the typical player experience in all regions. Much of the world has a 0 or near 0 Delta Ping. An important reference here is that a single server frame is 16ms. So the vast majority of players have a Delta Ping less than the duration of a single server frame. There are a few cases of large player count modes in relatively low search volume in regions (like the *Modern Warfare Zombies* mode in South Africa) where players are often matched out of their ideal region and DCs to play in other locations. We regularly see players in South Africa matching into European DCs to find matches for large player count modes. This data is used to inform our data center planning.

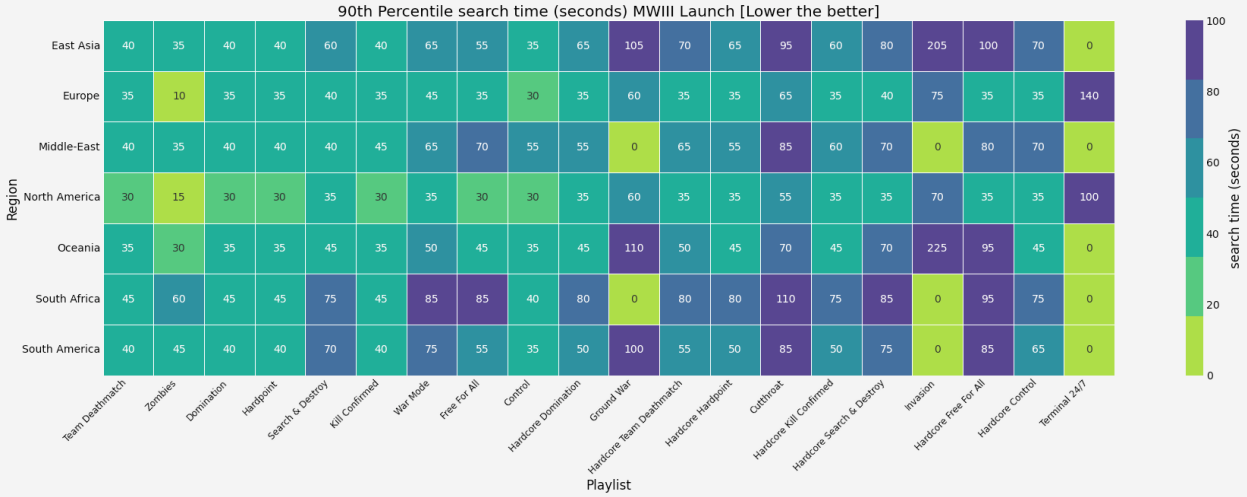
The below view is the 90th Percentile of Delta Ping. This covers the worst case scenario for players. These are typically searches at unusual times of day for the player's region. This greatly constrains our options, as we continue to pursue a short search time.




Here you can see that North American and European players are still consistently matched into low Delta Ping matches. EU is still around a single server frame in increased latency, while North America varies far more. This is due to the concentration of population centers on the coasts: as soon as it's necessary for a player on the East Coast to match into the West Coast, the ping will increase substantially. We see that here in the 90th Percentile.

Many other parts of the world, South Africa, East Asia, and some playlists in South America, don't have a sufficient local search volume to create matches in nearby DCs, during low traffic times of day. Ultimately, the best player experience for many of these players is found by matching them into other regions around the world. These regions being geographically distant from the high population regions force us to place them into higher latency games. The best example of this is perhaps Oceania, where the difference between playing locally and playing in a remote region is very clear. Most playlists even at the 90th percentile have a 0 delta ping (at the cost of search time), but modes like *Ground War* are pushed out of Oceania and into other regions (often North America). This pushes their Delta Ping up substantially.

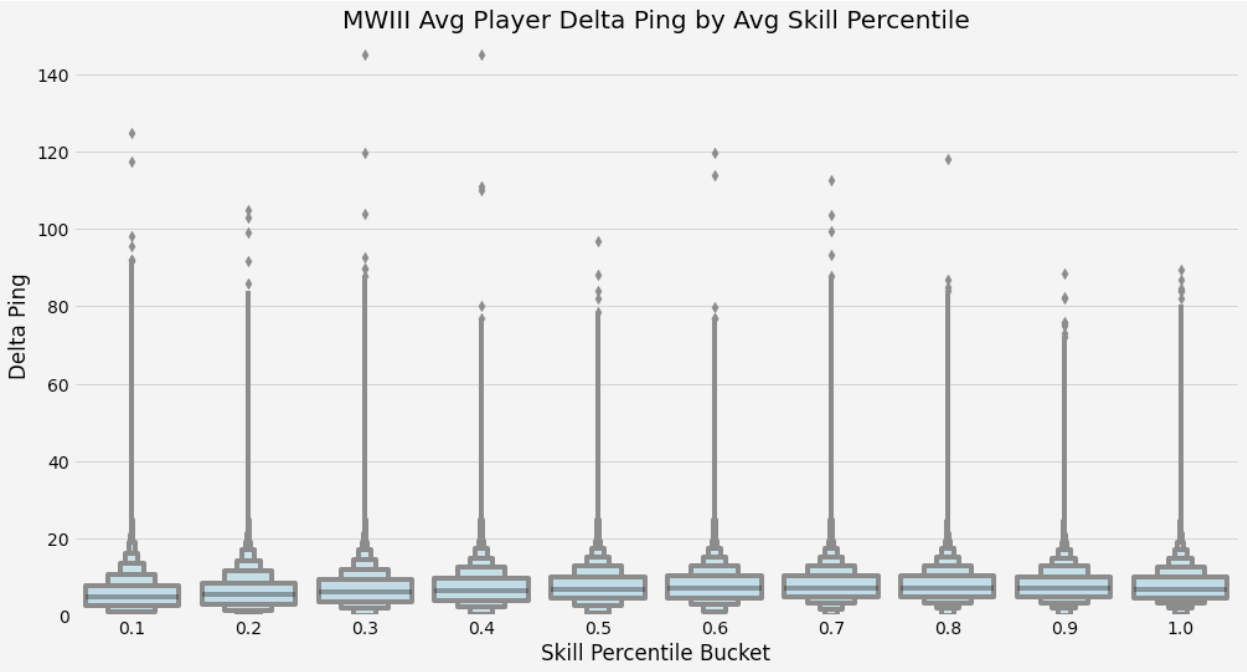
Comparing Oceania and North America is also a great example of how search time and ping trade off against one another. Below is a view of 90th percentile search time by region and playlist.




In North America, search times are kept 5-15 seconds faster than Oceania. We accomplish this by moving North American players (during these low search volume times of day) to nearby DCs coming at the cost of some Delta Ping. In Oceania, there aren't a lot of good alternatives in their region to host matches. As a result, their Delta ping stays low, but their search time increases. In the extreme cases, after we have failed to find matches in Oceania, we place them in matches in other regions, which comes at the cost of both Search Time and Delta Ping. This can be seen very clearly in the 90th percentile search time and delta ping for *Ground War* in Oceania.

All of this information is also used when considering future data center plans. Our goal is to minimize network latency for players. We use our monitoring data to look for sites that have good connectivity to a lot of players or that help close the gap between a large population center and a small one. By adding test sites to the list of data centers that clients ping we can assess the impact on the whole population and make better choices about where to add capacity. For example, our data center in Bahrain was added as it has good connectivity to most of the middle east as well as good connectivity to eastern Europe so it provides a link between those two regions.

Another important dimension to consider when evaluating our success in providing a low Delta Ping to all of our players, is skill. It’s important that no matter the skill level of the player, we provide a low latency experience. The visualization below breaks down our Dela Ping metric by skill decile. In it, you can see that our Delta Ping experience is highly consistent for all of our players, regardless of their skill level. In other words, a player's skill has no impact on the latency experience we provide.




# Summary

Reducing ping is a challenge for all online games and matchmaking systems and is particularly important for first person action games like *Call of Duty*.

*Call of Duty* uses a large number of data centers to provide local data centers where possible, and the matchmaking system uses a greedy optimization algorithm with an acceptable data center backoff to minimize the delta ping for players, subject to the limits of available search population.

The ping achieved by the matchmaking system is monitored to assess the effectiveness of our algorithms, and the suitability of the available data centers.

# References

[1] Activision Publishing. 2024. Call of Duty Update: An Inside Look at Matchmaking. Retrieved from https://www.callofduty.com/blog/2024/01/call-of-duty-update-an-Inside-look-at-matchmaking

[2] Paul Haile and Mitch Sanborn. 2019. Call of Duty: Modern Warfare Netcode Explained! Infinity Ward. Retrieved from https://www.youtube.com/watch?v=tCpYV4k_izE

[3] Blizzard Entertainment. 2019. Introducing Role Queue. Retrieved from https://overwatch.blizzard.com/en-us/news/23060961/introducing-role-queue/

[4] Joel Doonan and the FIFA Team. 2020. FIFA 20 Game Data Centers Deep Dives: New Location Process + Dallas FGDC. The Pitch Notes. Electronic Arts. Retrieved from https://www.ea.com/news/pitch-notes-fifa20-game-data-center-deep-dive