



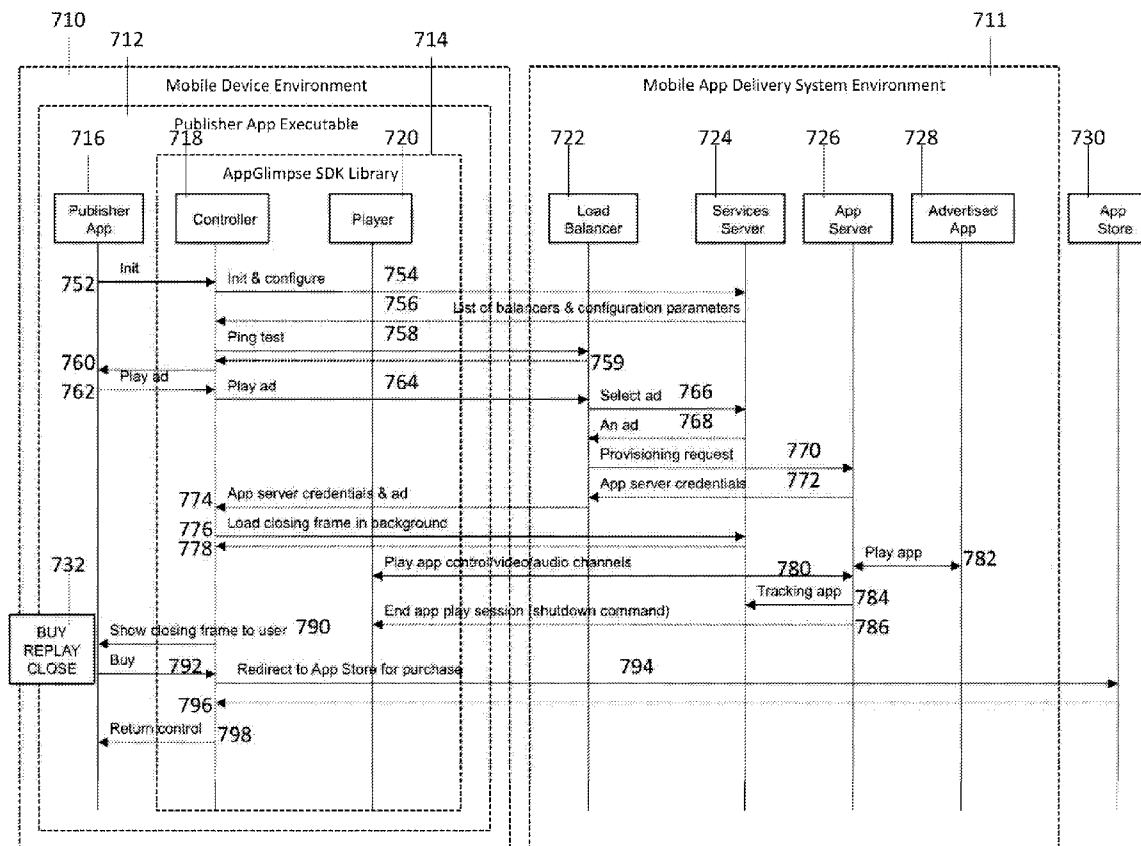
US 20150011311A1

(19) **United States**(12) **Patent Application Publication**
RELAN et al.(10) **Pub. No.: US 2015/0011311 A1**(43) **Pub. Date: Jan. 8, 2015**(54) **SYSTEM AND METHOD FOR
EXPERIENCING ADVERTISED
INTERACTIVE APPLICATIONS**(52) **U.S. Cl.**CPC *A63F 13/12* (2013.01); *A63F 2300/5506*
(2013.01)USPC **463/31**(71) Applicant: **ISWIFTER, Inc.**, Campbell, CA (US)(72) Inventors: **Rohan RELAN**, Atherton, CA (US);
Peter Kai-Hua TAN, Campbell, CA
(US); **Harsh NAYYAR**, Mountain View,
CA (US); **Ping-Hao WU**, San Francisco,
CA (US)(21) Appl. No.: **14/316,772**(22) Filed: **Jun. 26, 2014****Related U.S. Application Data**(60) Provisional application No. 61/842,958, filed on Jul. 3,
2013.**Publication Classification**(51) **Int. Cl.***A63F 13/30*

(2006.01)

(57) **ABSTRACT**

A client device executes a first interactive application that includes an embedded advertising player module. As the user interacts with the first interactive application, a condition is detected, causing the advertising player module to transmit a request for a second interactive application. The second interactive application may be developed to run in the same or different computing environment than the first interactive application. The second interactive application is executed by a remote server, and screen display outputs are generated in response. In response to the screen display outputs, the server generates video frames or rendering commands and streams the frames or commands to the advertising module to generate a display on the client device. The advertising player module receives interaction data for interacting with the second interactive application, and forwards such interaction data to the server. The second interactive application provides an output in response to the second interaction data.



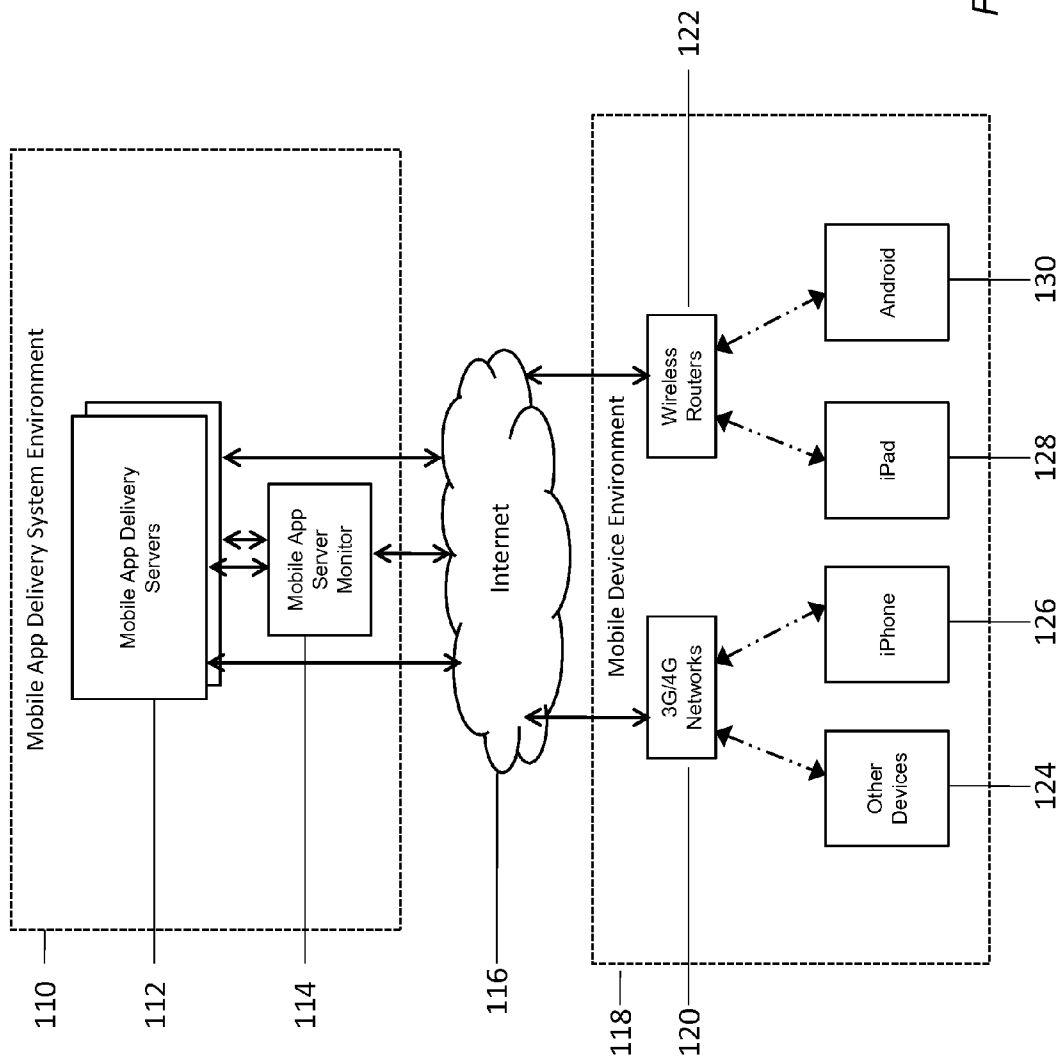
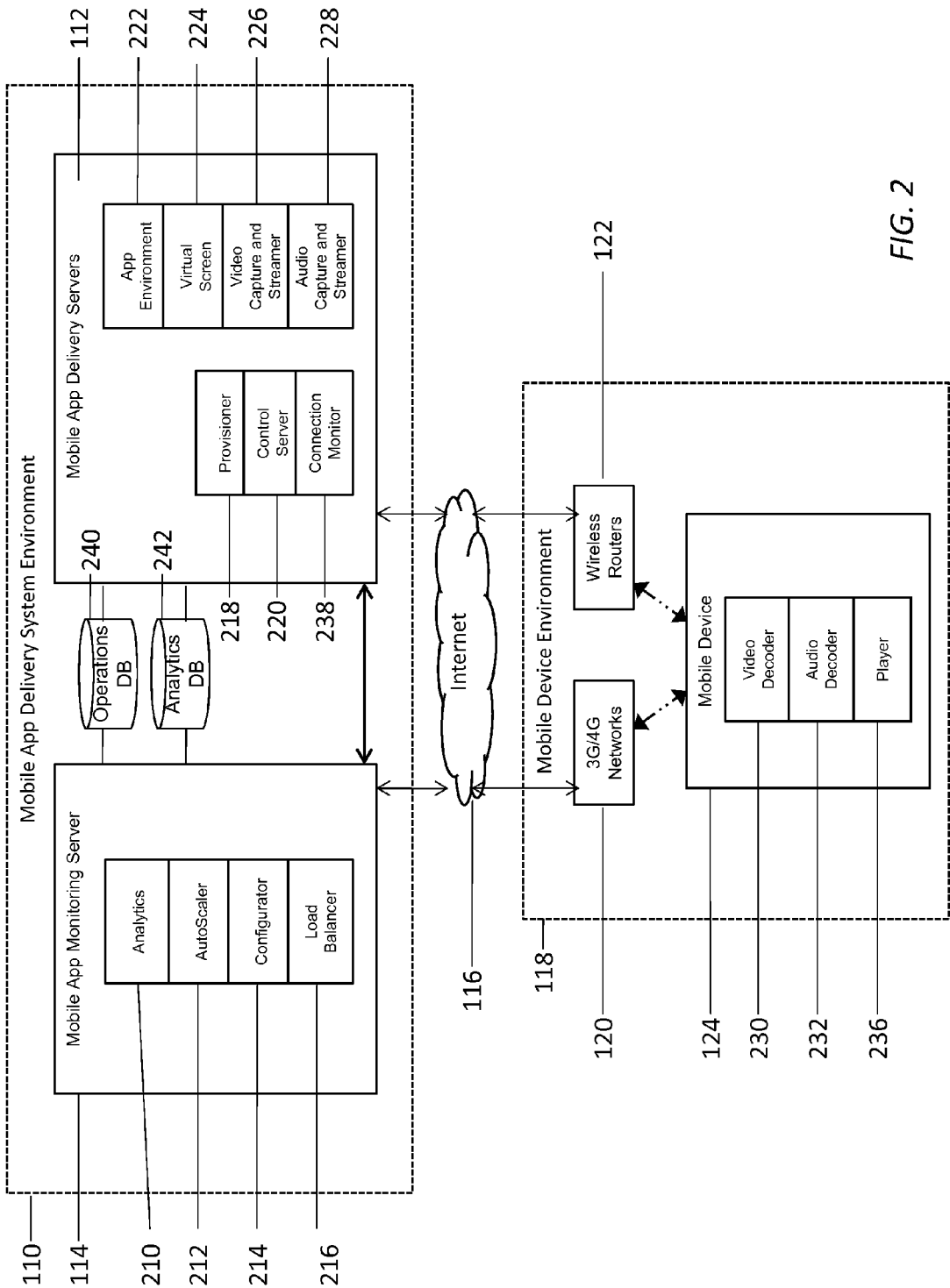


FIG. 1



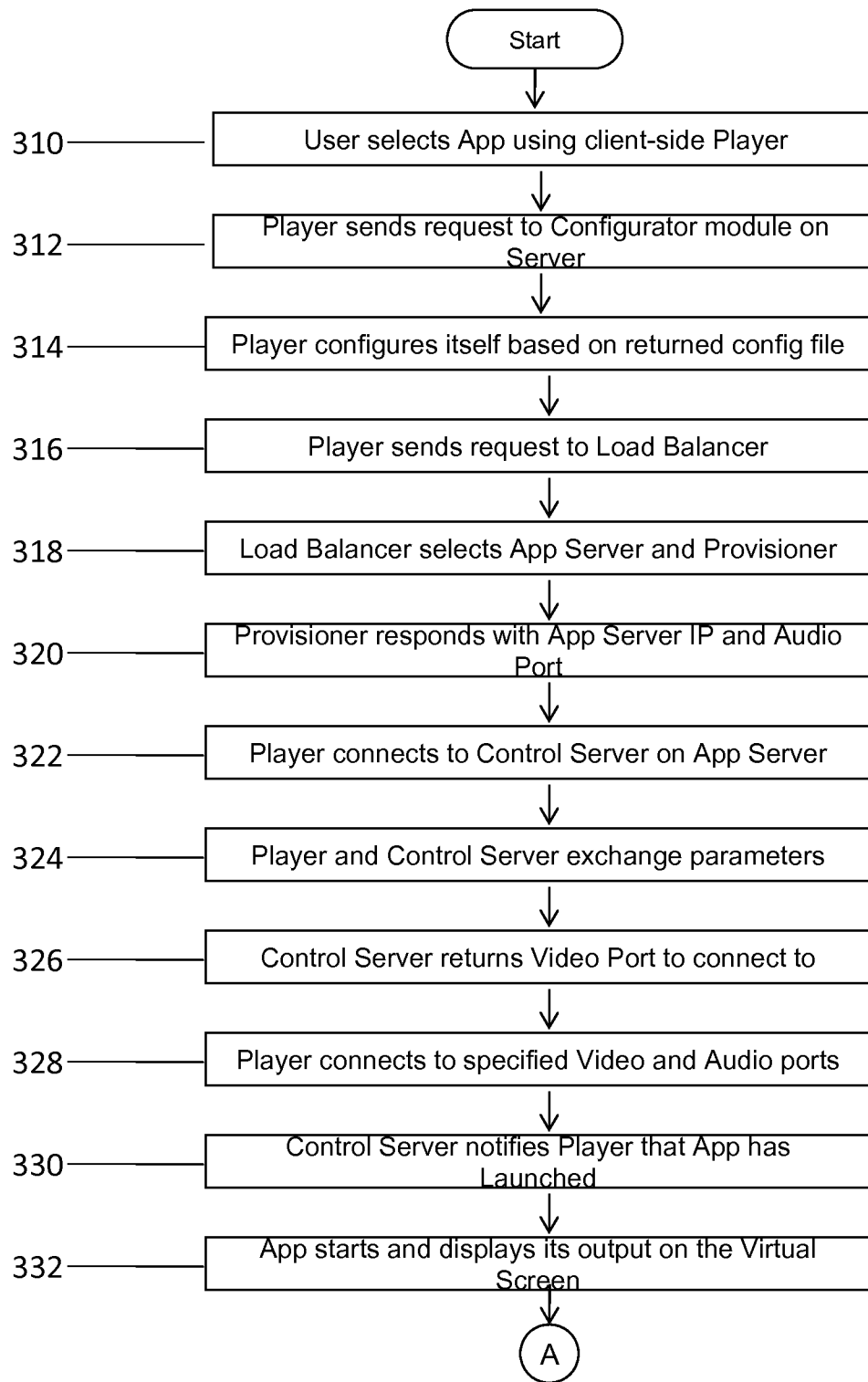


FIG. 3A

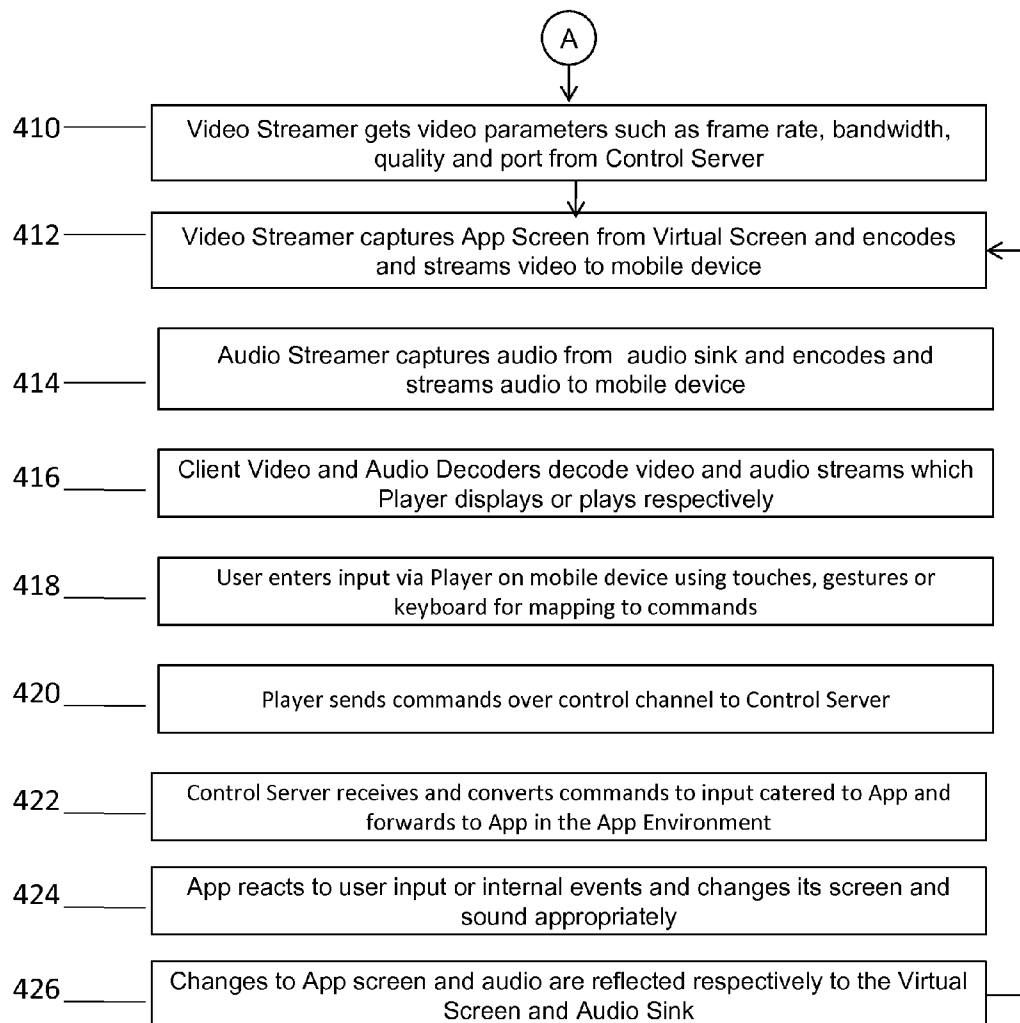


FIG. 3B

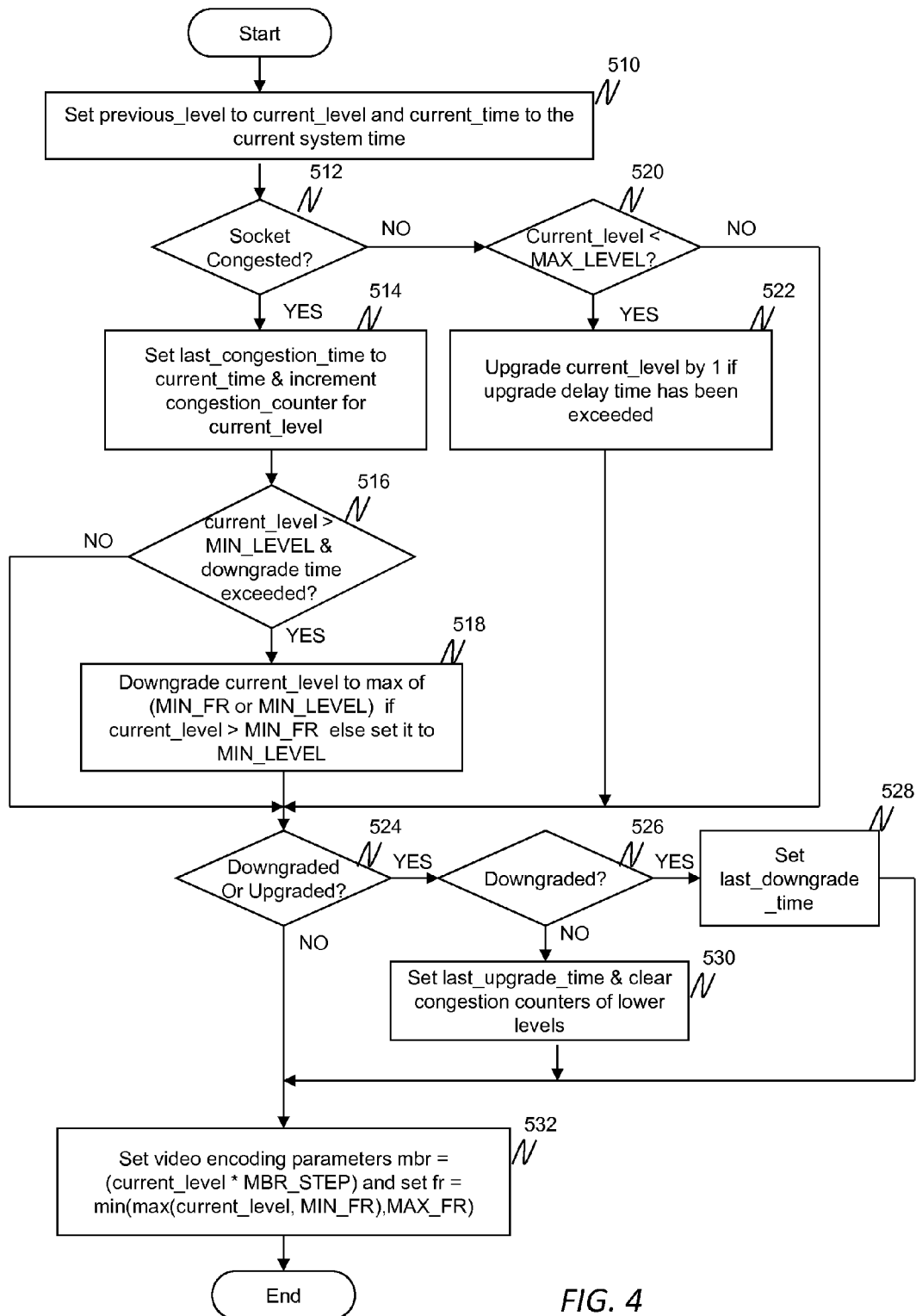
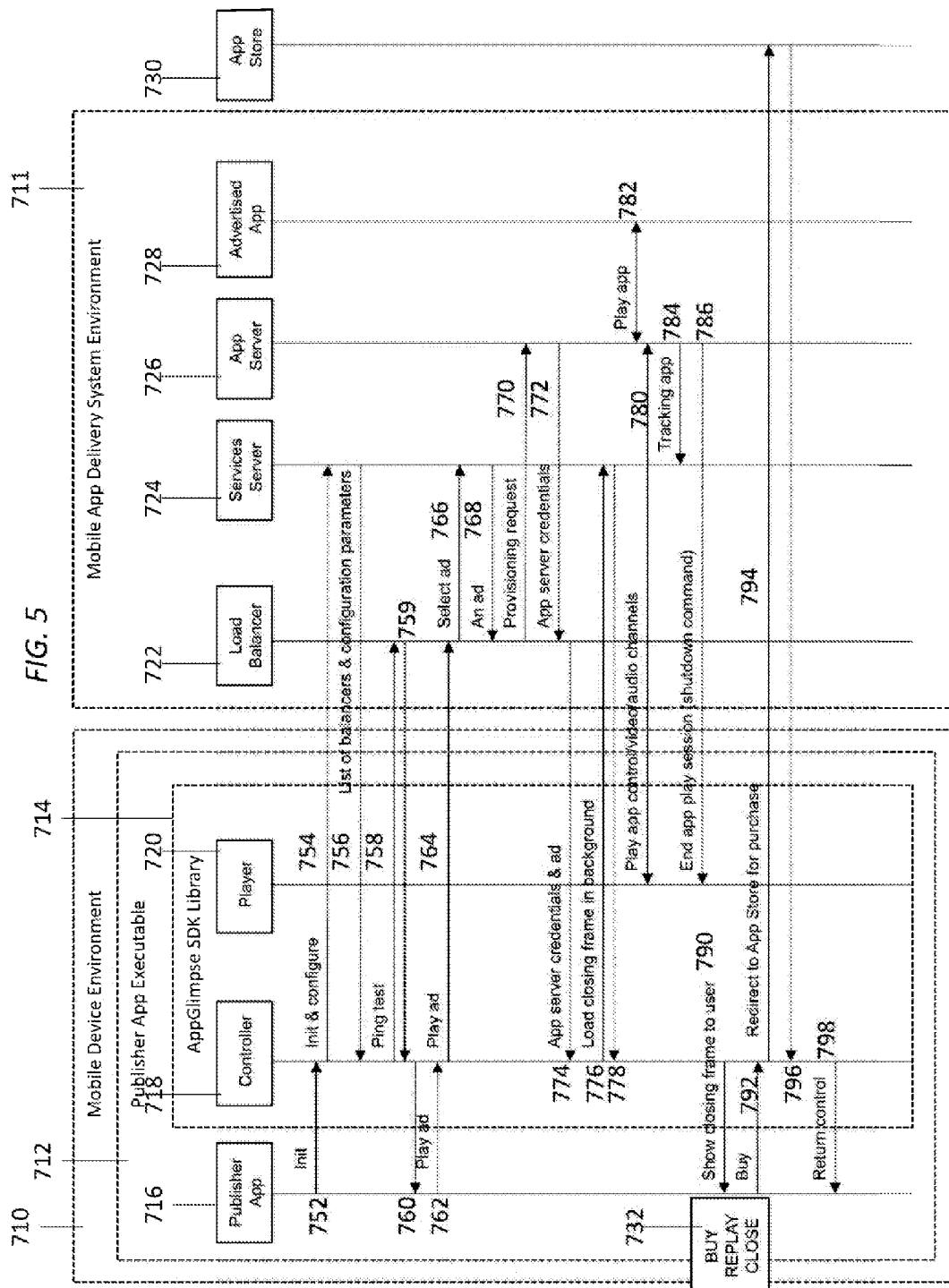


FIG. 4



SYSTEM AND METHOD FOR EXPERIENCING ADVERTISED INTERACTIVE APPLICATIONS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Patent Application Ser. No. 61/842,958 filed Jul. 3, 2013, the content of which is incorporated herein by reference.

FIELD OF THE INVENTION

[0002] The present invention relates generally to providing a software system and method by which software applications for mobile and non-mobile devices may be readily discovered, advertised, demoed, and purchased.

BACKGROUND

[0003] It is currently a problem for users to discover which software applications (hereinafter referred to as an “app” or “apps”) they might want to download or purchase for both mobile and non-mobile devices.

[0004] The top app stores (e.g. Apple App Store, Google Play Store, Amazon App Store) generally have tens or hundreds of thousands of apps, making it hard for developers to get their apps discovered by users once they have submitted their apps to these stores. App stores themselves tend to have limited “shelf space” where, generally speaking, only apps near the top of their respective lists (Top Paid, Top Free, Top Educational, etc.) get visibility, while the rest go unnoticed.

[0005] App store search capability is also quite anemic since developing page-rank-style algorithms for apps is difficult given the lack of semantic information about apps as compared to the web. App store searches are also often deficient in the most rudimentary search assistance one would expect on a web search. For example, whereas a web search engine may correct a user’s spelling or suggest alternate spellings for an item being searched, app store search engines often fail if the exact spelling for an app or category of app is not entered in the way in which the search engine indexes the app.

[0006] Accordingly, what is desired is a system and method for allowing apps to be discovered and experienced even if those apps are not readily visible in the app stores.

SUMMARY

[0007] Embodiments of the present invention are directed to a system and method for experiencing advertised interaction applications across a plurality of native and non-native app environments for both mobile and non-mobile device computing platforms. According to one embodiment, the method includes invoking, by a processor, a first interactive application where the first interactive application includes an embedded advertising player module. The processor receives first interaction data for interacting with the first interactive application. The processor invokes the advertising player module in response to the first interactive application detecting a condition. The condition may be, for example, reaching a particular goal in the first interactive application. The advertising player module transmits a request for a second interactive application. In response to the request, the second interactive application is invoked by a first server coupled to the processor. The second interactive application generates a screen display output, and in response to the screen display

output, video frames or rendering commands are streamed to the advertising module. The advertising player module receives the streamed video frames or rendering commands and generates a display in response. The advertising player module receives second interaction data for interacting with the second interactive application. The advertising player module forwards the second interaction data to the first server, and the second interactive application provides an output in response to the second interaction data.

[0008] According to one embodiment of the invention, the advertising player module displays an option relating to the second interactive application, receives a user input indicative of a selection of the option, and connects, in response to the user input, to a second server configured to generate an output responsive to the selected option.

[0009] According to one embodiment of the invention, the option is to purchase the second interactive application.

[0010] According to one embodiment of the invention, the advertising player module receives a message indicative of an end of the second interactive application. In response to receipt of the message, the advertising player module displays the second video frame including the option relating to the second interactive application.

[0011] According to one embodiment of the invention, an end of processing relating to the selected option is detected, and the processor invokes the first interactive application for resuming interaction with the first interactive application.

[0012] According to one embodiment of the invention, the native computing environment of the first interactive application is different from the native computing environment of the second interactive application.

[0013] According to one embodiment of the invention, each native computing environment provides an operating system, and the operating system in which the first interactive application is configured to run is different than the operating system in which the second interaction application is configured to run.

[0014] According to one embodiment of the invention, the processor receives the streamed video frames or rendering commands over a wide area network.

[0015] These and other features, aspects and advantages of the present invention will be more fully understood when considered with respect to the following detailed description, appended claims, and accompanying drawings. Of course, the actual scope of the invention is defined by the appended claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] FIG. 1 is an overview block diagram of a system to deliver applications to non-native client computing platforms according to one embodiment of the present invention;

[0017] FIG. 2 is a more detailed block diagram of various modules hosted by various computing device of FIG. 1 according to one embodiment of the present invention;

[0018] FIGS. 3A-3B are flow diagrams of a process for providing interactive content generated by an application configured to run in one application environment, to a client device providing a different application environment, according to one embodiment of the invention;

[0019] FIG. 4 is a flow diagram of a quality of service adaptation algorithm according to one embodiment of the invention; and

[0020] FIG. 5 is an overview block and flow diagram of a system and method by which apps for mobile and non-mobile devices may be readily discovered, advertised, experienced and purchased.

DETAILED DESCRIPTION

[0021] In an environment for developing, advertising and discovering apps, there are several role players: (1) a developer which is a person or entity that creates and develops an app; (2) an advertiser which is a person or entity that creates an ad unit for an app (sometimes the developer and advertiser are the same person or entity); (3) an ad network which is an entity that hosts ad units to be displayed and is an intermediary between advertisers and publishers; and (4) a publisher which is a person or entity that shows an ad unit to their users. A publisher is usually in fact also a developer of apps but publishes ad units within their own apps in order to monetize or gain revenue from their app. Developers/advertisers generally pay the ad networks to get their ads displayed. In turn, ad networks pay publishers to display streaming ad units to users.

[0022] Because the app store offers little discovery potential, developers often turn to advertising to reach users and get discovered. There are a few current approaches in advertising used by developers in getting their apps discovered.

[0023] One approach to app advertising is the use of banner ads. In this model, advertisers present a simple image in an attempt to entice users to click/tap on the image, which takes users to where they may download or buy the app. These banners may be presented on websites or on other native applications. One problem with this approach is that the ad does not offer much information to the user about the app, causing the users to be uninterested or misinformed.

[0024] Banner ads are also not very engaging. Thus, users do not have an incentive to complete the typically high friction download/purchasing process (e.g. tap on ad, redirect to app store, login to app store, download the app, launch the app, etc.). This friction often leads to low conversion rates on the ad, and users who are dissatisfied with the final product they download. Since click-through rates are low and the quality of the user experience is low, this type of low quality advertising often does not pay the ad publisher very well, often leading the publisher, advertiser, and user to all have an unsatisfying experience.

[0025] A second current approach to app advertising is using text banners, similar to banners ads, but with pure text instead of an image. The standard for this is adsense or adwords, where the text ad is displayed in the context of a webpage or search result. This means that the user is likely looking for a solution to the problem the developer's app may solve. However, text ads share the problems of banner advertising, as they do not provide a lot of information about the app before the user has to get into the high friction download/purchasing process. Text banners are also not very engaging.

[0026] A third current approach to app advertising is using video ads. Video ads are an attempt to solve the problem of banner and text ads by creating a more engaging ad unit. In this model, the advertiser creates a video showcasing the features of the app. The videos usually range from 15 seconds to a couple minutes. This has the benefit of providing more information to the user, leading the user to have a better understanding of what his final experience might be. The user is also generally more engaged, and therefore, more likely to follow through on completing the download/purchasing process.

Because users get a better idea of what the app does, they are more likely to be satisfied with the final app experience. This often leads to higher conversion and better user satisfaction. As a result, the ad publisher is also generally paid higher rates.

[0027] Video ads, however, are also not devoid of problems. Since video ads are not interactive, they are not as engaging as they could be, leading users to drop out before watching the whole video. Video ads may also not be completely representative of the final app experience, leading to users' dissatisfaction with their final experience after they complete the download/purchasing process. Video ads are also difficult for advertisers to create, as they are often expensive to produce, and use skill-sets and tools that are often very different from the skills and tools used to develop the app. Finally, app publishers may not like video ads because the publisher's apps in which the video ad units are embedded, are often fundamentally interactive experiences themselves, and watching a video is not interactive, which interrupts the user's flow of the publisher's app.

[0028] A fourth current approach to app advertising is the use of iAds (interactive ads). In this model, developers create simple interactions that users may perform with their ad. This might include, for example, tapping on an element in the ad unit to bring up more information on the subject of the advertisement, or rotating a view of an object. These units often need to be simplified because they exist in a simple execution environment where the number of interactions and outputs of these interactions are greatly reduced. Generating these types of ad units also generally requires additional work by the developer to operate in a new execution environment. These types of ads are also limited in their interactivity.

[0029] In a fifth current approach, an advertised app is bundled with a published app. In this alternative approach, the publisher bundles executable code and assets for the app-to-be-advertised into the publishing app and executes that code when it is ready to show the ad. This bundling approach has several problems: (1) it increases the size of the publisher app since it has to include code and possibly assets for all the apps it wants to advertise; (2) when the advertiser updates its app, all publishers of the app advertisement have to update their app if they want to include the new code; and (3) bundling one app into another this way often requires significant changes to the advertiser's app's code to enable it to run within another app, which is additional work for the advertiser.

[0030] In a sixth current approach, the advertised app code is downloaded instead of an ad unit. In this model, the publisher downloads the advertised app code and assets to be displayed, and executes the code when it is time to show the ad unit. This often eliminates issues 1 and 2 from the fifth approach, but still leaves issue 3 as described above. The sixth approach also has two additional issues associated with it. It generally causes a large download of app code and assets to demo the app, which uses up user's bandwidth and may cause a large delay before displaying the ad. Some computing platforms, such as iOS, do not allow the execution of downloaded code within another app, which prevents this solution from being used in those environments.

[0031] In a seventh current approach, the publishing app downloads the code only as needed and starts executing the code as soon as enough of it is ready to execute. This is typically called cloud paging. This partially resolves the issue of long download times by decreasing the time before the code may execute, but does not resolve the issue of some

computing platforms, like iOS, not allowing execution of downloaded code within another app. In addition, this approach continues to have the issue that bundling one app into another this way often requires significant changes to the advertiser's app code to enable it to run within another app, which is additional work for the advertiser. For some apps, the amount of code that needs to be downloaded before beginning execution may be very large which would also result in delay.

[0032] In an eight current approach, the advertiser rewrites its application in, for example, HTML+JavaScript (JS), which has execution environments available for downloaded code in iOS, Android, and the web. This way, the publisher may download the HTML/JS code when it wants to show the app and start executing it on most environments.

[0033] There are potential problems with this eighth approach. First, the approach could cause large downloads of code and assets, even if all those assets are not used. For example, the app may download an image that's used in the 50th second of the app demo, but the user closes the ad before getting there. This causes unnecessary bandwidth usage as well as possibly a delay before executing the ad. Second, the advertiser generally has to completely rewrite its application in a new language/environment, which is time consuming and expensive. This may require an entirely new codebase, which means new developers, tool-chains, APIs, etc. This is a huge hassle for developers and often not worth the effort. Third, the performance of HTML/JS code is generally not great, especially on mobile devices. For many applications, the performance is not sufficient to deliver a good user experience, so many advertisers do not even bother with this route since it will just provide a bad demo of their game and users will generally not convert.

[0034] Thus, there is a need for a system and method that provides a relatively easy and engaging way in which potential users may discover and experience apps they may be interested in, without having to first go through the high friction download/purchasing process. For the app developer, there is a need for a system and method that leverages the app developers existing skill-sets, tools, and apps, in order to provide potential users an experience of their app, without the need to produce expensive advertisements that are inherently not interactive. For the app publisher, there is a need for relatively easy and platform neutral way, that is also legal, by which they may integrate the ability to display any advertised app.

[0035] According to one embodiment of the invention, an ad unit, hereinafter called an AppGlimpse ad unit, is designed for interactive content, such as games and apps, which allows the user to interact with the app directly. An ad unit is a selected portion of an advertised app. Typically, the creator of an advertised app, selects a portion or portions of the advertised app to serve as the ad unit for that advertised app.

[0036] According to one embodiment, the playing of an AppGlimpse ad unit (for an advertised app) is initiated by a publisher app, native to a particular client computing platform, which has linked a library, hereinafter referred to as an AppGlimpse SDK library, to its executable. For example, when a user is playing a game like "Cannabalt" on a mobile device, like an iPad, and that game is also a publisher app, at the end of a game level the "Cannabalt" publisher app may choose to show the user an AppGlimpse ad unit.

[0037] In one example, the AppGlimpse ad unit may be a playable demo of "World of Goo", another game available for the iPad. According to one embodiment, the experience the

user gets playing "World of Goo" in the AppGlimpse ad unit may be very similar to the experience they would get playing the game after downloading and purchasing it, for example, on an app store. The user may play the game via a player embedded the publisher app, for a limited period of time, after which they are asked to perform some action if they wish to keep playing. This action may be the next step in the process (e.g. download or buy it from the app store).

[0038] In another embodiment of the invention, the user could download a specific publisher app from the app store (e.g. an application which allows a user to try an app before buying the app), which offers a number of such demos of games and other apps available on the app store. The user may try each app for a limited time before deciding whether they want to continue on to download and purchase the app from the app store. As noted, the publisher app therefore may be a first party app designed for the purpose of presenting AppGlimpse ad units using, for example, the AppGlimpse SDK library.

[0039] According to one embodiment of the present invention, the system and method for discovering, advertising, experiencing, demoing, downloading and purchasing apps across a plurality of native and non-native app environments for both mobile and non-mobile device platforms, includes both client-side and server-side components. In one embodiment of the present invention, the client-side device contains at least one publisher app native to a particular mobile or non-mobile platform (e.g. Android, iOS, AppleTV, and/or other computing platforms). The publisher app provides some function of its own as well as the ability to host AppGlimpse ad units for advertised apps. The publisher app may be configured to host the AppGlimpse ad units by linking an AppGlimpse SDK library to its executable.

[0040] According to one embodiment, the AppGlimpse SDK library includes the runtime support needed to control and play an AppGlimpse ad session. According to one embodiment, the AppGlimpse SDK library also includes an AppGlimpse player capable of playing streamed video, audio, OpenGL frames/commands, and the like, from an app streaming server.

[0041] According to one embodiment, the advertised app does not execute on the client device. Rather, the advertised app executes on the server-side in one or more app virtualization and streaming servers, and the advertised app's presentation is streamed as video, audio, or OpenGL frames/commands, to a player module (hereinafter referred to as a AppGlimpse player) on the client-device.

[0042] In one embodiment of the present invention, the server-side includes AppGlimpse ad units, which may be native to any computing platform supported by embodiments of the present invention. According to one embodiment, the AppGlimpse ad units are portions of advertised apps, selected by the advertiser, to be part of the ad units.

[0043] According to one embodiment of the present invention, the AppGlimpse ad unit of the advertised app executes within the app virtualization and app streaming servers as disclosed in further detail below.

[0044] According to one embodiment of the present invention, the advertiser/developer of the advertised app provides a version of their app that they would like streamed as the AppGlimpse ad unit. This version of their app may be in one of many forms, including but not limited to a PC executable, a Linux executable, Android APK, an iOS app, or the like.

[0045] According to one embodiment of the invention, the virtual execution and rendering environment described in further detail below, converts the rendering screens of the app into video/audio frames or OpenGL commands which are streamed down to the client device instantly and in real-time using the AppGlimpse player of the SDK library at the client side for presentation. Also, according to one embodiment, the client-side AppGlimpse player converts input received via an input interface from the user (touches, clicks, and the like) into packets on the network, which are streamed up using a low latency channel, and converted into inputs for the advertised app running in the app virtualization environment.

[0046] According to one embodiment of the invention, publishers are provided with an AppGlimpse SDK library, which the publisher may integrate/embed into their publisher apps. According to one embodiment, the AppGlimpse SDK library provides the code necessary to control and play an AppGlimpse ad unit session. The SDKs features include but are not limited to: 1) being able to choose when to display the ad; 2) being able to decide on the size of the ad; 3) being able to theme the ad; 4) being able to provide analytics sent from the ad; 5) being able to choose the types of ads to show; 6) being able to determine the frequency to show the ad; and 6) being able to specify the amount of incentivized goods to deliver to the user after successful completion of the ad.

[0047] On the client-side, according to one embodiment of the invention, the interface between the publisher app and the AppGlimpse SDK library includes but is not limited to two primary interface calls. The first interface is an initialization call referred to as an “init()” call. When “init” is called by the publisher app, the controller component of the AppGlimpse SDK library on the client makes an “init” request over a network to mobile app delivery system servers. A services server responds to this request by returning a list of load balancers plus various initialization parameters, including but not limited to bandwidth, quality-of-service, frame-rate, bit-rate, stereo/mono, and the like. The controller component on the client performs a ping test to each of the load balancers in the list. According to one embodiment, the controller component selects the load balancer with the shortest ping time response. According to one embodiment, it is this load balancer that will be used subsequently when a request to play an ad is received.

[0048] According to one embodiment of the invention, the second primary interface is a play call referred to as a “play()” call. When the publisher app is ready to play an AppGlimpse ad unit, perhaps after a user finishes a certain level in a game, the publisher app calls “play()”. The controller component in the AppGlimpse SDK library responds to the play command by sending a request over the network to the previously selected load balancer requesting the ID of an AppGlimpse ad unit to be played. On the server-side, the load balancer, upon receiving the play command, calls the services server requesting an ad selection. The services server returns an ad selection to the load balancer. The load balancer, using a load balance algorithm, calls a particular app/game server with a provisioning request. The app/game server returns its server credentials (e.g. IP address) to the load balancer. The load balancer then returns both the server credentials and the AppGlimpse ad unit ID back to the controller component inside the AppGlimpse SDK library on the client. According to one embodiment, the controller makes a background call across the network to the services server requesting a closing frame to be displayed after the AppGlimpse ad unit session is

over. According to one embodiment, the closing frame contains the choices available to the user after experiencing the AppGlimpse ad unit, such as, for example, “Replay”, “Buy”, “Close,” and the like. After the request for a closing frame is initiated, the controller component of the AppGlimpse SDK library gives control to the AppGlimpse player in the library, passing it the server credentials including an AppGlimpse ad unit ID.

[0049] According to one embodiment of the present invention, the AppGlimpse player starts an app streaming session with the app/game server on the server side. According to one embodiment, the app/game server on the server-side executes the selected advertised app in its app virtualization environment and streams video, audio, or OpenGL frames or commands, to the AppGlimpse player on the client. The AppGlimpse player decodes the streamed data and presents the data to the user. The AppGlimpse player also collects all input (e.g. user interaction data), and forwards the input data to the app/game server over a control channel over the network. In addition, as the advertised app is streamed to the client, the app/game server reports at various points a tracking progress to the services server. Once the AppGlimpse ad unit is completed, the app/game server sends a shutdown command over the network to the controller component of the AppGlimpse SDK library on the client. The controller component then shows the closing frame to the user (e.g. “Buy,” “Replay,” “Close,” and/or the like), and responds to the user’s input. According to one embodiment, if the user selects “Buy,” the user is redirected to the appropriate app store to begin the purchase download process for the advertised app. At the end of processing for the closing frame, the controller returns control to the publisher app.

[0050] As a person of skill in the art should recognize, an AppGlimpse ad unit is not limited to being displayed on mobile devices. The AppGlimpse ad unit may also be displayed on a web browser as well as on other computing platforms such as, for example, a Smart TV.

[0051] According to one embodiment, the AppGlimpse ad unit is not limited to the computing platform for which the original app was developed. For example an app developed exclusively for Android phones may be discovered, advertised, and experienced on any other computing platform such as, for example, an iPhone, a web browser, Apple TV, and/or any other mobile or non-mobile computing platforms. Also, an AppGlimpse ad unit may be displayed in many different forms, including but not limited to, pre-rolls, post-rolls, interstitials, expanding banners, and incentivized play.

[0052] As a person of skill in the art should appreciate, the app discovery approach according to embodiments of the present invention has several advantages over previous approaches. For the user, it offers a more engaging ad unit since the user may actually interact with the ad unit in a meaningful way. It also offers a very good representation of what the final app experience will be like, so the user is better informed if they decide to download the app as to what sort of experience they will have. The AppGlimpse ad units also load very quickly compared to prior art mechanisms, so the user is not waiting needlessly. The process of experiencing the AppGlimpse ad unit is frictionless for the user. There are no extra steps required such as redirecting to the app store, log-ins, downloads or anything else associated with the traditional process of discovering apps on app stores.

[0053] For the advertiser, one benefit is that the user is more engaged with their ad unit, leading to better branding and

higher conversion rates through the download/purchase process. The advertiser is also likely to have more satisfied users since the users have a good idea of what they are getting prior to purchase/download.

[0054] The AppGlimpse ad unit is also not very difficult for the developer/advertiser to create, as, according to one embodiment, it just uses a modified version of the original app, which the developer has the skills and tools in-house to produce. In some cases where the original app is also the ad unit, it may take no work at all. The advertiser may also get analytics about how the AppGlimpse ad unit is performing and at what points the user engagement drops off, allowing the advertiser to optimize their final app/game-play experience as well.

[0055] For the makers of publisher apps, the higher conversions result in higher rates paid to the publisher. The makers of publisher apps may also prefer the more interactive ad experience provided by the AppGlimpse ad unit, because their own publishing app is fundamentally interactive.

[0056] According to one embodiment of the invention, advertisers also receive a dashboard which includes but is not limited to: analytics on engagement at various points in the ad; frequency capping; where they would like to show their ad, ability to bid/pay for each install, and the like.

[0057] Because, in one embodiment of the present invention, the present invention leverages app streaming technology to deliver interactive ad units, it provides the publisher the ability to provide a fully interactive experience of any portion of their original app to a potential buyer on any mobile or non-mobile device platform, including even those platforms that are not native to the original app, resulting in a full range of interactivity and flexibility to the publisher.

[0058] Embodiments of the present invention allow for a highly interactive and engaging ad network/demo system with the following benefits:

[0059] (1) the system does not require a lot of work from the creator of the advertised app because the creator may leverage his existing app code within the virtualization environment of an embodiment of the present invention;

[0060] (2) it is easy for the publisher app to integrate a virtualization player into the publisher app.;

[0061] (3) the system does not have a huge performance penalty;

[0062] (4) the system does not require upfront downloading of a lot of code/assets;

[0063] (5) the system works in environments that do not allow the executable code of one app to be embedded into another app (such as the iOS platform);

[0064] (6) the system works across all client environments because of the flexibility of the remote streaming of video, audio, and commands according to embodiments of the present invention.

I. System and Method for Providing Interactive Content

[0065] Embodiments of the present invention provide a system and method for delivering interactive content generated by software applications configured to run on a native application environment, to clients providing non-native application environments, over a data communications network. Such applications may be, for example, gaming applications, and the clients may be mobile phones. The non-native application environments provided by the client devices do not provide the same APIs for running software applications that are provided by the native application envi-

ronments in which the software applications are intended to run. Although embodiments of the present invention are directed to providing the interactive content to non-native application environments over the data communications network, a person of skill in the art should recognize that the interactive content may also be provided to native platforms.

[0066] In general terms, when the client is a mobile device, the method for delivering the interactive content includes:

[0067] 1. Selecting on the mobile device a server-resident application to be presented on the mobile device.

[0068] 2. Customization of how the application will appear on the mobile device.

[0069] 3. Initialization of control, video and audio channels to be used to deliver application content to the mobile device.

[0070] 4. Selection of a server to run the application.

[0071] 5. Startup of the application on the selected server.

[0072] 6. Server-side screen capture and audio capture of the output of the selected application.

[0073] 7. Server-side encoding and streaming of the captured video and audio to the client device during a streaming session.

[0074] 8. Client-side decoding and display of the streamed video on the mobile device.

[0075] 9. Client-side decoding and playing of the streamed audio on the client device.

[0076] 10. Client-side input acceptance and encoding on the client device.

[0077] 11. Client-side forwarding of input to the server-side non-native application.

[0078] 12. Converting the input to a corresponding input catered to the particular application.

[0079] 13. Continuous monitoring of the condition of the data communications network during the streaming session and dynamically adjusting the video streams for achieving a highly responsive and interactive experience for the user of the non-native application on the mobile device.

[0080] FIG. 1 is an overview block diagram of a system for efficiently delivering interactive content to non-native client devices and platforms according to one embodiment of the present invention. The system in FIG. 1 includes a mobile application delivery system environment 110, a mobile device environment 118, and a data communications network 116 interconnecting the two environments. The data communications network 116 may be a local area network, private wide area network, or the public Internet, accessible via any wired or wireless technology conventional in the art. In one embodiment of the invention, the mobile application delivery system environment 110 is a cloud computing web server environment.

[0081] According to one embodiment of the invention, the mobile application delivery system environment 110 includes a set of delivery servers (also referred to as host devices) 112 and a monitoring server 114. According to one embodiment, each of the delivery servers host a software application on a native application environment. The native application environment provides, at a minimum, the same APIs as the APIs available on the original application environment for which the software application was specifically developed to run.

[0082] The monitoring server 114 takes requests to start a new application and selects a specific server from a pool of delivery servers 112 that host the application. According to one embodiment, the delivery servers 112 and/or monitoring server 114 are deployed and run in different geographic locations spread out from one another. In order to give the users a

highly responsive interactive experience, the software components that reside in client devices **124** send ICMP echo requests to a load balancer in a monitoring server **114** in a different geographic location, measure the roundtrip time, and choose the delivery server with lowest latency.

[0083] According to one embodiment of the invention, the mobile device environment **118** includes standard carrier 3G/4G networks **120**, wireless routers **122**, and various client devices **124-130** (collectively referenced as **124**). The client devices may be mobile phones, electronic tablets, laptops, TV set top boxes, smart TVs, and the like, having a processor, memory, input interface (e.g. interface to a keyboard, mouse, joystick, microphone, etc.), output interface (e.g. interface to speakers, display screen, etc.), and network interface for connecting to a data communications network. Although the environment **118** is described as a mobile environment, a person of skill in the art should recognize that the environment may include other environments, as such, for example, wired environments that include wired devices.

[0084] Each of the delivery servers **112**, monitoring server **114**, and client devices **124** includes a central processing unit (CPU) for executing software instructions and interacting with other system components for performing the functions described herein. The servers and client devices further include a mass storage device such as, for example, a hard disk drive or drive array, for storing various applications and data used for implementing the system. The server and client devices further include an addressable memory for storing software instructions to be executed by the CPU.

[0085] The server and client devices further include various input and output units conventional in the art. For example, each device may have an input unit such as, for example, a keyboard, keypad, joystick, microphone, and/or display screens with pressure sensitive (touch screen) technology. Each device may also have an output unit such as, for example, speakers, display screens, and the like. The server and client devices may further include wired or wireless data communication links for accessing the data communications network **116**.

[0086] FIG. 2 is an overview block diagram of various modules hosted by the monitoring servers **114**, delivery servers **112**, and mobile devices **124** according to one embodiment of the present invention. The various modules are implemented via computer program instructions which are stored in memory for executing by the CPU of the corresponding server. A person of skill in the art should recognize, however, that all or a portion of the various modules may be implemented via firmware, hardware, or a combination of software, firmware, and/or hardware.

[0087] According to one embodiment of the invention, the modules that reside in the monitoring servers **114** include, but are not limited to, a load balancer **216**, a configurator **214** module, an autoscaler **212** module, and an analytics **210** module.

[0088] The load balancer **216** is configured to find a delivery server **112** that may support an incoming connection request to start an application. The load balancer is configured to select a machine and process instance for each new application request. Once the load balancer selects a machine instance, it also selects the instance of a control server, provisioner, and application environment modules that will be used for a particular instance of the application.

[0089] The load balancer **216** is configured to maximize the number of users on each delivery server **112**. This allows

more delivery servers **112** to become idle, i.e. without any active connections, so the autoscaler **212** may shut them down, potentially saving on cost.

[0090] According to one embodiment of the invention the load balancer **216** uses a connection monitor **238** on each delivery server **112** to find an available server. The connection monitor **238** may be implemented as a Ruby process that polls every second for a number of active connections on its delivery server **112**. This includes polling for the number of active application environment processes, such as Firefox processes, Wine processes, or custom processes. The connection monitor **238** sends the collected data to an operations database **240**. In one embodiment of the present invention, the operations database **240** is high performance database such as, for example, a Mongodb database, configured to handle a high amount of inserts/updates per second as well as quickly respond to queries. According to one embodiment, the connection monitor **238** sends its hostname, the number of open connections, and the number of available connections (the maximum available connections minus the open connections on a machine).

[0091] When a request comes in to the load balancer **216**, it queries the operations database **240** to find a delivery server **112** that has connections available. Since there is a lag between when a client has been allocated to a particular delivery server **112**, and when the client connects and the connections monitor **238** records the connection, the load balancer **216** cannot entirely trust the data coming from the operations database **240**. For example, if a delivery server **112** is near capacity but has one connection available, and the load balancer **216** receives several requests in a short time frame, it may forward all requests to the same delivery server **112** before any client connects and increases the connection count.

[0092] In one embodiment of the invention, to mitigate the risk of this happening, two strategies are implemented. First, the load balancer **216** is configured to randomly select from a pool of candidate delivery servers **112** so requests do not always go to the same delivery server **112**. Second, the load balancer **216** is configured to keep track of the last time it sent a request to a particular delivery server **112** so as to not send the same server multiple requests in a short time period. If there are no delivery servers **112** with available connections, the load balancer **216** is configured to try a preset number of times (e.g. three times) to find one before simply allocating the connection to a random server. This helps ensure that every user always receives a connection, even though his or her experience may be slightly degraded by the excess load on the server.

[0093] In one embodiment of the present invention, the load balancer **216** is further configured to automatically remove problematic delivery servers **112** by checking the responses from the servers. If the response is an error, the server is queued for deletion. Similarly, if the response takes longer than a threshold time period or the delivery server **112** is unreachable, it is queued for deletion.

[0094] According to one embodiment of the invention, the autoscaler **212** is a module configured to manage the delivery servers **112** in order to accomplish various tasks. For example, the autoscaler may be configured to make sure enough delivery servers **112** are available to serve current users as well as users that may be connecting in the near future. The autoscaler may also be configured to delete unneeded delivery servers **112** so that there is not excess

capacity, thus potentially reducing costs, such as, for example, in an embodiment that uses cloud servers.

[0095] In this regard, the autoscaler **212** regularly polls the operations database **240** and checks the number of available connections. If the number of available connections is too low, the autoscaler **212** starts enough delivery servers **112** to a configuration specified target number. If the number of available connections is too high, the autoscaler **212** queues the oldest delivery servers **112** for deletion until the number of available connections is equal to the required buffer size. Queuing the oldest servers for deletion helps reduce the number of errors in the system, as typically the oldest servers are prone to errors. A delivery server **112** that is queued for deletion is not immediately deleted as it may have users currently active. However, a server queued for deletion does not receive new connections. The autoscaler **212** regularly checks if any deleted delivery servers **112** are idle and deletes them when they are.

[0096] In one embodiment of the invention the autoscaler **212** is implemented as a set of Ruby processes. The autoscaler **212** may also interact with the cloud providers, such as Blue Box Group and Amazon Web Services, for creating and deleting servers.

[0097] According to one embodiment of the invention, the analytics module **210** keeps statistics on system operations. Analytic data is stored in the analytics database **242**. The analytics database **242** may be implemented as a high performance database such as, for example, a MongoDB database. The load balancer **216** stores information about user requests in the analytics database **242**. The control server **220** updates records in the analytics database **242** when a user of a client device logs out so that total session time may be calculated. The analytics module **210** is designed to take analytic data offline for further processing. In addition, statistics pertaining to application requests and session distribution are available to be queried from the monitoring server **114**.

[0098] According to one embodiment of the invention, the configurator **214** is a module configured to select and return a configuration file when a user selects a software application or game (collectively referred to as an application) on the client. The configuration files allow customization on the client on a per application basis for elements like mapping user gestures or taps to the original inputs of the application (e.g. mouse down, mouse click events, and the like). The configuration file also supports customization of “player skins” in order to customize the look and feel of the display on the client device.

[0099] According to one embodiment, configuration files store information on how the player **236** on the client device **124** should interact with the system, request video frame rate parameters, load content, and the like. There are two types of configuration files: player configuration files and application or game specific configuration files. The player **236** on the client device **124** consumes player configuration files. These files primarily have the URLs of services for retrieval by the player **236**. For example a list of apps/games to show on the player **236**; user ratings of apps/games, update purchase state in the case of a purchase or trial expiration, and the like. Some parameters may be used to distinguish between different client devices, such as, for example, between an electronic tablet and a mobile phone.

[0100] Application or game specific configuration files are used when a particular application or game is being started. This configuration file has all application or game specific

information, like the URL to load, the product ID for payment, the screen dimensions it will run at, and parameters to distinguish between the various client devices (e.g. between iPad and an iPhone).

[0101] In one embodiment of the invention, configuration files are stored as JSON files. The configuration files allow the player **236** to be as general as possible, and thus allows changes to content or behavior on the fly via changes of the configuration files from the delivery server **112**.

[0102] According to one embodiment of the invention, the provisioner **218** is a module on the delivery servers **112** which sets up a user's account on a particular delivery server. In one embodiment of the present invention, each unique user and its corresponding connection runs in its own Linux user account on the delivery server **112**. This allows the sandboxing of users from each other. It also allows the creation of a secure system, as the user accounts have stripped permissions and are only allowed access to files and processes that are required for the selected application. The provisioner **218** creates the user account and adds the user to the required group.

[0103] In one embodiment of the present invention, the user account set up by the provisioner **218** is determined based on the requested application. The skeleton directory has the files required for the particular application that is selected, for example the correct Mozilla file for a Flash application or the correct executable for a Wine application. The provisioner **218** also creates files in the user's directory to set parameters for the user. Such parameters may include, for example, the audio port to be used, Facebook credentials needed, and the like. The provisioner **218** is configured to set up some of the services the user may require. For instance, in one embodiment of the present invention, the provisioner **218** creates a pulseaudio sink for the user needed to stream audio to the client device.

[0104] According to one embodiment of the invention, the application environment **222** is a computer environment in which an application executes. The application environment **222** is specific to the type of application selected by the client device, and a computing platform (or virtualized computing platform based on a different underlying platform) of the delivery servers **112**. According to one embodiment, an application environment includes, but is not limited to frameworks, libraries, APIs, and services for the runtime execution of programs developed for a particular computing platform. The application environment may also provide a virtualization layer to provide a different or isolated computing platform. For example, when the particular operating system provided by the computing platform of the delivery servers **112** is Linux, and the selected application is a browser-based application or a Flash-based application, the application environment **222** may be a Firefox browser or other similar browser configured to execute the Flash-based application. If the selected application is a Microsoft Windows application, the application environment **222** is Wine or other similar environment which allows computer programs written for Microsoft Windows to run on Unix-like operating systems. If the selected application requires a graphics processing unit (GPU), such as, for example, for high quality Windows games, the application environment **222** is vglrun. Vglrun executes the application with VirtualGL, which provides OpenGL to Xvnc. When native Unix or Java applications are selected, they are simply run as native applications and the application environment **222** is Unix or Linux.

[0105] When the application environment 222 is the Firefox browser, certain Firefox extensions are also used to provide additional functionality and interactivity for a user. For example, a Greasemonkey extension and scripts, and a custom extension, are configured to be used to achieve actions like scrolling, toggling the visibility of the keyboard, URL logging, and the like. The Greasemonkey extension is used to add or edit an HTML DOM that is loaded. The extension associates a script, written in JavaScript, to a URL, and the extension is triggered only for the associated URL. The URLs are filtered through regular expressions. The Greasemonkey scripts trigger after the associated HTML page has been loaded, which serves as an indication that the desired content is loaded, and signals the control server 220 to send a message to the client device 124 to remove the loading screen. The Greasemonkey scripts also attach event handlers to input elements on the HTML page that instruct the control server 220 to show or hide the keyboard. The Greasemonkey scripts are also used to weed out unwanted elements like advertisements and therefore only have the desired content appear on a virtual screen 224 for being streamed to the client devices 124-130.

[0106] According to one embodiment of the invention, the Greasemonkey scripts are served up for use remotely through a web service. There is one script present in the Greasemonkey extension that is triggered for an URL that is loaded. The Greasemonkey script then queries a web service passing in the current URL and the content type as parameters. The web service looks up a configuration file that in one embodiment of the present invention is encoded in JSON. The web server then serves the corresponding script back to the Greasemonkey script for it to execute. The configuration file has information of the URLs against which a particular script is to be triggered. The configuration file also has information of a list of excluded URLs and content types for which scripts are not to be triggered. All the URLs are filtered through regular expressions.

[0107] According to one embodiment of the invention, when the application environment 222 is the Firefox browser, a customized Redirector extension is also used. When the Firefox browser is launched and the customized extension is initialized, the Redirector extension creates a socket to hard-coded port on the control server 220. The customized extension performs actions such as page scrolling, page refreshing, stopping page loads, blocking URLs, and the like.

[0108] According to one embodiment, the customized extension is used to perform page commands forwarded from the client device 124. For example, when the player 236 on the client device 124, in response to a user action sends commands such as scroll page, refresh page, stop page, load URL, and the like, to the control server 220 on the delivery server 112, the control server 220 relays these commands to the customized extension. The customized extension in turn parses the commands and performs the appropriate action as a script on the Firefox browser. According to one embodiment, the customized extension may also be used to block URLs. The Firefox browser asks the customized extension if it should load a URL, and depending on the category of the URL, the request is served or cancelled. The list of blocked URLs comes from the configuration file of the content that has been currently loaded. The URL request is filtered through regular expressions. There are five categories of URLs:

[0109] 1. White listed URLs: when such a URL is requested, the extension serves the }}request.

[0110] 2. Black listed URLs: when such a URL is requested, the extension cancels the request.

[0111] 3. Pop UP listed URLs: when such a URL is requested to load, the extension cancels the request and signals the client device 124, via the control server 220 and the player 236 on the client device, to show a message to the user informing them that the pop up was blocked.

[0112] 4. Browser listed URLs: when such a URL is requested to load, the extension cancels the request and signals the client device 124, via the control server 220 and the player 236 on the client device, to show a message to the user to load the URL natively.

[0113] 5. Ad URLs: when a URL does not match any of the above URLs, the extension cancels the request and signals the client device 124, via the control server 220 and the player 236 on the client device, to show the URL loaded in a web view inside of the player.

[0114] The customized extension for the Firefox browser application environment may also signal the client appropriately with a message to purchase an application after a trial period is over. When the customized extension is initialized, it queries a service for the payment information of the user for the content loaded. If the content is paid for, the user is given an uninterrupted browsing session. If the user has not yet paid for the content, he/she is considered a trial user. When the trial is over, the extension signals the client device 124, via the control server 220 and the player 236 on the mobile device, to display a message that the trial period is over and provide an option to buy the content. Any further requests are cancelled until the user purchases the content. Using a similar mechanism during the trial period, the user is presented, at periodic intervals, with messages in pop-up windows to make a purchase. In one embodiment of the invention, this payment setup is used for subscription payment systems.

[0115] In yet another example, if the native computing environment for the application that is hosted is Android, the application environment 222 module is a virtualization/emulation software and Android OS. The virtualization/emulation software provides a computing environment for the Android OS to run. To instrument the application, a custom launcher is used to control the launching of applications. A launcher is a main view of the OS and is responsible for starting other applications. The default launcher of Android is a home screen that a user usually sees. The custom launcher enables launching into the application directly when a client connects, and also prevents the application from exiting when user presses the back button. To achieve a seamless user experience, the data specific to mobile devices 124 are also captured in the mobile device and sent through the control channel to the control server 220. This data may include the device orientation, GPS location, gyro/accelerometer data, and the like.

[0116] According to one embodiment of the invention, the control server 220 is a module that authenticates a client based on input username and password. The control server 220 is also configured to receive input commands from the user via the player 236 on the mobile device 124 and forward these commands to the application environment 222 for the current application. The control server 220 is further configured to forward commands and messages from the application environment 222 to the player 236 on the client device 124 to control the look of the screen on the client device, such

as for example, to display notifications that an application has launched, that a pop-up should be displayed, and the like.

[0117] In one embodiment of the present invention, authentication is done using PAM authentication. According to one embodiment, the provisioner **218** creates the usernames and passwords for each user. Once the control server **220** has authenticated the user, it knows what username to use for subsequent operations.

[0118] According to embodiment of the invention, once the control server **220** has authenticated the mobile client, it executes two processes. The first process is executed by the virtual screen **224** module. According to one embodiment, the virtual screen **224** module is an Xvnc session which provides a X11 display for the application the user wants to run. Xvnc is the Unix VNC (Virtual Network Computing) server, which is based on a standard X server. According to this embodiment, applications may display themselves on Xvnc as if it were a normal X display, but the applications are configured to actually appear on any connected VNC viewers rather than on a physical screen.

[0119] The control server **220** may also execute a bash script that sets any preferences needed for the application selected, such as Firefox preferences. The script starts the audio and video streamers, the window manager, and the application the user requested.

[0120] The second process or environment the control server **220** starts is the application environment **222** for the selected application. According to one embodiment of the invention, the control server **220** maintains a description of which specific application environments are used based on the type of application selected and the environment of the delivery server **112**.

[0121] According to one embodiment of the invention, the video capture and streamer **226** module is a module that captures video frames output by the virtual screen **224** module. After the frames are captured, the video capture and streamer **226** module subsequently encodes and streams the video to the mobile device **124**.

[0122] In one embodiment of the present invention, the video capture and streamer **226** encodes in near real-time without a large buffer of frames and streams the encoded frames to the mobile device **124**, where the video decoder **230** on the mobile device **124** decodes in near real-time, also without a large buffer of frames, thus achieving a highly responsive video display.

[0123] In one embodiment of the present invention, at start-up, the video streamer and capture **226** module connects to the control server **220**. The control server **220** sends the video capture and streamer **226** module the parameters it should use, including frame rate, bit rate, and a video port to stream the video to the mobile device **124**. The connection between the control server **220** and the video capture and streamer **226** is maintained throughout the lifetime of the connection from the client device **124**, and may be used for quality of service adjustments of the video stream.

[0124] According to one embodiment, video capturing and encoding parameters are selected in a way that the system fully utilizes the network bandwidth designated to the video stream to produce high-quality video while keeping the encoding complexity sufficiently low to minimize the encoding time required for each frame, as part of the low-latency design. To further achieve minimal latency in the system so that applications are highly responsive to user input, the server encodes each video frame right after being captured,

and the output from the encoder is immediately fed into the socket without excessive buffering or delay. On the client, the video decoder continuously parses the incoming video data from the socket. As soon as it collects all the data required to decode a video frame, the frame is decoded, resized if needed, converted to an appropriate color space, and displayed. Again, video data are decoded and displayed as soon as they are available and there is no excessive buffering to ensure minimal latency.

[0125] According to one embodiment of the invention, the video capture and streamer **226** module consists of a set of scripts (e.g. Python scripts) for capturing, encoding, and streaming video to the mobile device **124**. According to one embodiment, the script launches one or more video processing programs (e.g. programs implemented using video processing libraries provided in FFMPEG) for the capturing, encoding, and streaming.

[0126] According to one embodiment, the video capture and streamer **226** module captures the display output by the virtual screen **224** module. For example, if using FFMPEG, the x11 grab function is invoked to capture the display from the Xvnc display. The video capture and streamer **226** then encodes (e.g. using FFMPEG together with x264) the video according to preset parameters. According to one embodiment, the various parameters are configured to provide low-latency operations to achieve real-time responsiveness to inputs provided by a user. The video capture and streamer **226** module captures the encoded output and streams the video to the mobile device **124**. The video stream connection to the mobile device **124** is maintained for the lifetime of the mobile device **124** client connection. The mobile device **124** decodes the video using the video decoder **230** module, which in one embodiment of the invention, uses the H.264 decoder from the avcodec library in FFMPEG. The resulting frames are displayed by the player **236** module on the client device **124**. The screen display may be done using a low-level API such as, for example, OpenGL ES (e.g. on iOS-based mobile devices), Surfaces (e.g. for Android-based mobile devices), and the like.

[0127] In one embodiment of the present invention, the video is encoded by the video capture and streamer **226** module on the delivery server **112** into a byte stream (e.g. an H.264 Annex B byte-stream using FFMPEG and x264), and streamed to the client device **124** through a TCP socket. The video capturing and encoding is based on a video frame rate (in frames/sec or fps) and maximum video bit rate (in bits/sec or bps). The two parameters together determine a maximum frame size which identifies a maximum number of bits that the encoder may spend to encode a frame. The resulting visual quality of the frame is affected based on the maximum frame size that is selected.

[0128] On the client device, the video decoder **230** module parses the byte-stream into data units each corresponding to an encoded video frame and feeds them sequentially a decoder (e.g. H.264 decoder). However, because of how H.264 Annex B byte-streams are defined, the video decoder **230** module is able to recognize the completion of an encoded frame once the start of the next frame is observed. Therefore, a video frame cannot be correctly parsed and decoded until the next frame is received, resulting in an extra latency of the duration of one frame in the display of video content, significant for an interactive system. According to one embodiment, in order to address this latency, the video capture and streamer **226** module is configured to prefix a 4-byte field for each

encoded video frame in the byte-stream, indicating the number of bytes included in the encoded frame. With this prefix, the video decoder **230** on the client device **124** may extract an encoded frame from the byte-stream as soon as it arrives and pass it to the decoder without having to wait for the next frame.

[0129] According to another embodiment, the video capture and streamer **226** module is configured to append a tag for marking the end of a video frame. For example, an AUD (Access unit Delimiter) defined in H.264 may be appended to the end of each encoded frame. Although AUD in H.264 is designed for marking the beginning of a video frame, it is inserted into the stream right after an encoded frame instead, without having to wait for the next frame to be captured or encoded. The decoder on the client detects the AUD immediately after receiving an encoded frame and starts decoding the frame without having to wait for the next frame.

[0130] In one embodiment of the present invention, the video capture, encoding, and/or streaming parameters may be adjusted in real-time to allow for quality, latency, and/or bandwidth trade-offs, depending on the particular application or game. For example, in one embodiment of the invention, the lowest latency settings are used for Flash-based applications to provide better responsiveness at the expense of video quality. However, for certain Window's games, higher latency settings are used to avoid screen pixilation. These parameters are initially set in the configuration file for a particular application or game, and are configured to be modified in real-time. For example, parameters such as the video frame rate and/or maximum video bit rate may be adjusted in real-time while streaming a particular application or game, based on a monitored condition of the network.

[0131] More specifically, embodiments of the present invention provide two solutions to stream video from the delivery servers **112** to the client devices **124** with low latency so that applications provide real-time response to user input, while maintaining good visual quality: a first solution based on TCP (Transmission Control Protocol); and a second solution based on UDP (User Datagram Protocol). In both solutions, the virtual screen display on the server is periodically captured into video frames based on the frame rate specified in the system. The video frames are encoded and delivered from the video capture and streamer **226** on the server to the video decoder **230** on the client, via either a TCP or an UDP socket. Video capturing and encoding parameters are selected in a way that the system fully utilizes the network bandwidth designated to the video stream to produce high-quality video while keeping the encoding complexity sufficiently low to minimize the encoding time required for each frame, as part of the low-latency design. To further achieve minimal latency in the system, on the server each video frame is encoded right after being captured, and the output from the encoder is immediately fed into the socket without excessive buffering or delay. On the client, the video decoder continuously parses the incoming video data from the socket. As soon as it collects all the data required to decode a video frame, the frame is decoded, resized if needed and converted to an appropriate color space, and displayed. Again, video data are decoded and displayed as soon as they are available and there is no excessive buffering to ensure minimal latency.

[0132] TCP-Based Solution

[0133] According to one embodiment, the TCP-based solution uses a TCP socket to provide a reliable channel for video delivery. Potential transmission errors that may occur due to,

for example, temporary glitches in the network or insufficient network bandwidth are all taken care of by the built-in error-detection and retransmission mechanism in TCP. Typical video streaming systems insert I-frames periodically in the video stream to enable random access and error recovery capability in the video decoder. However, for high-quality videos the I-frames are usually difficult to compress and therefore may take more time to transmit, resulting in a surge in latency. In the TCP-based solution, since error recovery is no longer needed at the decoder level and random access capability is not required in the interactive streaming application, except for the first frame in the video stream, I-frames are not used to avoid such a latency surge.

[0134] When transmission errors occur in the network, the TCP socket on the server side automatically slows down outgoing transmission, and the video traffic coming into the socket from the video streamer may be congested and buffered in the socket, increasing the video latency experienced on the client. To relieve such congestion, the amount of video traffic sent into the socket is quickly decreased so that the latency may be reduced back to a minimal level. In this regard, a QoS (Quality of Service) adaptation algorithm dynamically adjusts the video capturing and encoding parameters based on the network condition. According to one embodiment, the QoS adaptation algorithm not only downgrades video QoS, hence decreasing the amount of video traffic during bad network conditions to reduce latency, but also upgrades video QoS when additional network bandwidth is available to further improve the video quality. The QoS adaptation algorithm determines two QoS parameters used in video capturing and encoding: video: (1) frame rate (in frames/sec or fps); and (2) maximum video bitrate (in bits/sec or bps), denoted by fr and mbr respectively. The two parameters, fr and mbr, together determine the maximum frame size (mbrlfr), the max. number of bits that the video encoder may spend to encode a frame, which helps provide a good indication of the resulting visual quality of the frame. The operational range of the QoS parameters is defined by four system constants MIN_FR, MAX_FR, MIN_MBR and MAX_MBR such that $MIN_FR \leq fr \leq MAX_FR$ and $MIN_MBR \leq mbr \leq MAX_MBR$. Another constant MBR_STEP (in bits/sec or bps) is specified in the algorithm to map the QoS parameters into a discrete set of QoS levels, together with two time intervals DOWNGRADE_INTERVAL and UPGRADE_INTERVAL that define how much time the algorithm has to wait before making a QoS adjustment. Table 1 is a pseudo code of the QoS algorithm for the TCP-based solution.

TABLE 1

```

define MAX_LEVEL as floor(MAX_MBR/MBR_STEP)
define MIN_LEVEL as ceil(MIN_MBR/MBR_STEP)
last_downgrade_time = 0
last_upgrade_time = 0
last_congestion_time = 0
for each level that MIN_LEVEL <= level <= MAX_LEVEL
    congestion_counter[level] = 0
end
current_level = min(max(MAX_FR, MIN_LEVEL), MAX_LEVEL)
for each video frame
    previous_level = current_level
    current_time = current system time
    if (socket is congested in the previous transmission)
        last_congestion_time = current_time
        congestion_counter[current_level] =
            congestion_counter[current_level] + 1

```

TABLE 1-continued

```

    if (current_level > MIN_LEVEL)
        if (current_time - last_downgrade_time >
            DOWNGRADE_INTERVAL)
            if (current_level > MIN_FR)
                current_level = max(MIN_FR, MIN_LEVEL)
            else
                current_level = MIN_LEVEL
            end
        end
    end
else
    if (current_level < MAX_LEVEL)
        multiplier = pow(2, congestion_counter[current_level +
            1])
        if (current_time - max(last_upgrade_time,
            last_congestion_time) >
            multiplier * UPGRADE_INTERVAL)
            current_level = current_level + 1
        end
    end
end
if (current_level < previous_level)
    last_downgrade_time = current_time
else if (current_level > previous_level)
    last_upgrade_time = current_time
    for each level that MIN_LEVEL <= level <= previous_level
        congestion_counter[level] = 0
    end
end
mbr = current_level * MBR_STEP
fr = min(max(current_level, MIN_FR), MAX_FR)
end

```

[0135] According to one embodiment, a congestion in the TCP socket is detected by testing if new data may be written into the socket within a certain timeout time (e.g., 20 msec), using, for example, a linux poll command. Upon congestion in the socket, the video capture and streamer 226 drops the video frame rate to a minimum value and correspondingly reduces the maximum video bit rate. In addition, when congestion occurs at a certain QoS level, the congestion count for that level is increased to record the congestion. If no congestion has occurred for a preset time since the last congestion or last QoS upgrade, indicating a good network condition, the algorithm gradually upgrades the QoS parameters. This interval increases, exponentially for example, with the congestion count at the destination QoS level, making it more difficult to upgrade to a level already shown to be congestion-prone. If the algorithm stays at a certain QoS level without a congestion long enough that it upgrades to the next level, the original level proves to be well supported by the network and the congestion counts of the level and all levels below it are reset to zero. According to one embodiment, when the value of the QoS level becomes smaller than MIN_FR, the video framerate fr stays at MIN_FR but the visual quality of each frame further degrades as indicated by the decrease in the max. frame size mbrlfr. Similarly, when the QoS level becomes larger than MAX_FR, fr stays at MAX_FR but quality of each frame is further improved.

[0136] In general, a video encoder does not always produce video traffic at the specified maximum bit rate. The video traffic generated may be far below mbr when the video content is rather static, and only approaches mbr during fast-moving scenes. This leads to an issue that congestion may never occur in a static scene, and the QoS adaptation algorithm keeps upgrading the QoS level multiple times although the upgraded mbr is already well above the bandwidth supported by the network. Significant congestion, hence high

latency, may then be set off by a sudden scene change in the video content that generates traffic approaching the highly overvalued mbr. According to one embodiment, to resolve this issue, during a certain time interval right after a QoS upgrade, the video encoder is modified to append dummy data that is ignored in the decoder at the end of each encoded frame so that the frame size reaches the maximum frame size mbr/fr. This modification helps ensure that the video traffic temporarily reaches the specified mbr right after each QoS upgrade. If this bitrate is already more than the network may support, a minor congestion with less impact on latency is triggered followed by a timely QoS downgrade.

[0137] UDP-Based Solution

[0138] According to one embodiment, a retransmission mechanism in the TCP-based solution helps guarantee that all the data fed into the socket by the video streamer eventually arrive at the video decoder. However, during a bad network condition, the additional time required for retransmission and the potential congestion in the socket may have an adverse impact on the video latency. Although the QoS adaptation algorithm is configured to quickly relieve the congestion, the transient latency increase usually manifests in the form of a momentary pause in the video, which is undesirable in some applications. The UDP-based solution tackles the problem differently. The UDP-based solution does not attempt any retransmission. Data is not buffered in the socket and does not get discarded if not delivered in time. Therefore, video latency is less influenced by the network condition and stays rather constant, at the expense of potential transmission errors at the decoder level. In terms of user experience, the UDP-based solution differs from the TCP-based solution in that instead of potentially introducing momentary video pauses, occasional decoding errors may become noticeable.

[0139] The UDP-based solution uses RTP (Real-time Transport Protocol) to deliver the video stream from the server to the client over a UDP socket. RTP handles reordering of the UDP packets and detects if any packet gets lost in the transmission. Without retransmission, a lost packet may lead to decoding errors visible in the decoded frame that cannot be fully removed by error concealment techniques in the decoder. These visible decoding errors may also propagate across multiple frames because of the inter-frame dependency in the video stream. To stop such propagation, whenever a lost packet is detected a request for an intra-refresh cycle is sent from the client to the server through the control channel. Instead of using I-frames that may create a surge in latency, intra refresh is a technique in video coding that evenly distributes intra blocks that may be decoded independently from previous frames over multiple frames to stop error propagation while minimizing the increase in video traffic.

[0140] In a typical network setup, UDP packets get lost occasionally even when the available network bandwidth is sufficient for the video traffic. However, when packet losses occur frequently, an assumption is made that the current network bandwidth may no longer support the video traffic, and the video QoS level is downgraded. In this regard, a QoS adaptation algorithm similar to the one for the TCP-based solution is used for UDP with two main differences. First, the algorithm for UDP runs on the client instead of on the server. The QoS level determined on the client is then signaled to the server through the control channel to adjust the parameters in video capturing and encoding. Second, the algorithm detects if the frequency of UDP packet losses exceeds a threshold,

rather than testing for congestion in the TCP socket, to decide if a QoS downgrade is needed or if an upgrade may be performed.

[0141] According to one embodiment of the invention, the audio capture and streamer 228 module is a module that captures audio produced by the application. After the audio is captured the audio capture and streamer 228 module subsequently encodes and streams audio to the client device 124.

[0142] In one embodiment of the present invention, audio is captured by the audio capture and streamer 228 module using pulseaudio. Each application has a pulse sink created for it by the provisioner 218. When the application is launched, the pulse sink for the application is set to the one created by the provisioner 218. This ensures that all audio from the application goes to the pulse sink for the user requesting the application. For non pulse-aware applications, a wrapper script, such as padsp, is used to provide an OSS to pulseaudio bridge.

[0143] In one embodiment of the present invention, to capture the audio, parec is used with the source set to be the sink's monitor. The output from parec is passed into an audio encoder.

[0144] In one embodiment of the present invention, the audio encoder may be either CELT, MP3 or the audio may be sent uncompressed. The setting used is determined by what the client supports.

[0145] In one embodiment of the present invention, the output from the audio encoder is passed to the UDP audio streamer. Upon receiving the audio stream, the client device 124, using the audio decoder 232 module decodes the audio. In one embodiment of the invention the audio decoder 232 uses the CELT library. In another embodiment of the invention, the audio decoder 232 uses the native iPhone MP3 decoding. The audio decoder 232 in the client device 124 uses a ring buffer to ensure too much audio data doesn't buffer up and create latency. It also uses Audio Queue Service for low latency operation.

[0146] In one embodiment of the present invention, the audio capture and streamer 228, uses Python scripts to monitor all audio processes on the delivery server 112. If any process dies, all the audio processes are restarted and the user only experiences a brief and sometimes unnoticeable audio interruption.

[0147] According to one embodiment of the invention, the player 236 is a module that plays on the client device 124, the application that was selected by the user and which is executed on the delivery servers 112. According to one embodiment, the player allows a user of the client device to play content provided by a software application that is not originally intended to run in the application environment provided by the client device. For example, the client device does not provide the necessary APIs provided by the native application environment of the software application. The client device may thus be described as providing a non-native application environment with respect to the software application. In this regard, the software application is run on the delivery server 112, and the player 236 displays video and plays audio output of the application that is streamed from the delivery server to the client device 124. The player 236 also accepts input from the user, maps the input to a predefined command, encodes the command using a binary protocol, and sends the encoded command over the control channel to the delivery server 112. The mapping of the user input to predefined commands is based on a configuration file and/or modified based on the content currently displayed.

[0148] The player 236 communicates with the server side modules using a control channel. According to one embodiment, all communication between the client and server, which the exception of streamed video and audio, occurs over the control channel. Separate video and audio ports on the server are used to stream the video and audio.

[0149] According to one embodiment of the invention, the video decoder 230 decodes video streamed from the server. In one embodiment of the invention, the video decoder 230 uses the H.264 decoder from the avcodec library in FFMPEG.

[0150] According to one embodiment of the invention, the video decoder 230 decodes the incoming video stream into video frames in a YUV420 format. To playback the video frames with a desired color representation and dimension, the video decoder converts the video frames back to a RGB color space and resizes the video frames before being displayed on the screen. According to one embodiment, the color-space conversion and resizing is performed using programming shaders running on a graphics processing unit (GPU) through graphics APIs (e.g., OpenGL) that are generally available on clients such as iOS or Android devices. Color-space conversion and resizing using GPUs rather than CPUs significantly reduces the processing time due generally to the parallel processing architecture of GPUs especially suitable for such tasks, resulting in a reduced latency in the system as well as a higher frame rate that may be supported on the client.

[0151] According to one embodiment of the invention, the player 236 is independent of and not specific to any particular application content that it will present on the client device 124. Only one version of the player 236 is needed for a particular client device 124, regardless of how many different applications or application types it presents on the client device.

[0152] According to one embodiment of the invention, all communications between the player 236 on the client device 124 and the control server 220 on the delivery server 112 happens over the control channel. The player 236 takes input in the form of touches or on-screen keyboard selections. The input may also be provided via other input devices such as, for example, physical keypads, joysticks, and the like. The input is mapped to preset command and sent from the player 236 over the control channel to the control server 220. The control server 220 converts the received command into a command applicable for the application (e.g. mouse-down command, mouse-up command, or keyboard command) using a simulation tool such as, for example, libxdotool. Thus, the control server is configured to translate an input provided using a particular input device, to a second input that is generally provided using a different input device.

[0153] According to one embodiment of the invention, the player 236 may also send commands in response to user gestures. For example, a swipe down gesture is mapped to a scroll down command for a browser session. The mapping of the gesture to a particular command is based on the configuration file provided to the player for the particular application or game that is being played. The scroll down command is sent by the player 236 as a scroll down command over the control channel to the control server 220. The control server 220 forwards the scroll down command over a socket connection to the application environment 222. The application environment may run specific scripts to execute the particular command. For example, for Flash applications, a Greasemonkey script in the application environment 222 executes an input command using Javascript.

[0154] According to one embodiment of the invention, the player 236 and the control server 220, using custom commands communicated over the control channel, may coordinate the velocity of the user's scrolling on the client device 124 and the resulting scrolling of the video screen. For example, in one embodiment of the invention, the player 236 interprets the velocity of the user's finger movement into a scroll command of x pixels. The player 236 sends a custom command for a scroll of x pixels via the control channel to the control server 220, which in turn provides the command to the application environment 222. The application scrolls its screen by x pixels and outputs the updated screen for capture and streaming to the client device 124. The video decoder 230 decodes the updated video frames and the player 236 displays a screen where its contents have moved x pixels in response to the user's scrolling gesture. The rapid fine-grain scrolling on the client device and other user input which are communicated to the server causes the returned video and audio streams to reflect the user's scrolling and other action as if playing the application on its native device. This coordination of user input gestures on the client device 124, with reaction of the application on the delivery server 112, provides a highly interactive and responsive experience for the user.

[0155] According to one embodiment, the control channel on the client exposes many different types of actions which map to the different types of content. These are encoded using a binary protocol and sent to the server. The server executes these actions differently based on the type of content currently active. Below are a few examples:

[0156] 1) Control channel exposes: sendMouseDown(button, x-position, y-position)

[0157] Encoding: 2 byte integer indicating the message size (5 bytes), 1 byte indicating the button, 2 bytes for x position, 2 bytes for position

[0158] Executed: via a library that talks to the X11 protocol called xdotool

[0159] 2) Control channel exposes: sendOrientation-Changed(orientation)

[0160] Encoding: 2 byte integer indicating the message size (1 byte), 1 byte indicating the 4 possible orientation

[0161] Executed: via a TCP socket connected to the simulator that tells it to change orientation

[0162] 3) Control channel exposes: sendTouches(action, active_touch_id, number_touches, touch_id1, x-position1, y-position1, . . .)

[0163] Encoding: 2 byte integer indicating the message size (3+5*number_touches bytes), 1 byte indicating the type of action (touch down/move/up), 1 byte indicating the active touch ID that triggered this event, 1 byte the number of touches, followed by the information of each individual touch [touch_id (1 byte), x-position (2 byte), y-position (2 byte)].

[0164] Executed: via a TCP socket connected to the simulator that tells it the current touch event and all the touch data

[0165] On the client, each of these commands may be executed by different types of input. For example, sendMouseDown may be called when the "tap" gesture is recognized if the current server-side content is flash. However, if the current server-side content is an Android game, the sendTouches command may be executed in response to the "tap" gesture as this makes more sense.

[0166] According to one embodiment of the invention, instead of mapping a user input to a specific command catered

for the application at the client side, the raw input data is provided to the client and the server converts the raw input data to an input that is appropriate based on the application environment and/or context of the application. For example, if the server-side application environment is Firefox and the content inside is a web page, a touch down immediately followed by a series of touch moves and touch up (a swipe/drag action) on the client mobile device implies the user wants to scroll the page. However, if the control server is not aware that the content is a web page, the server sends a mouse down followed by mouse moves and a mouse up, which is a selection action in Firefox. In order to interpret a swipe/drag action (touch down->move->up) as a scroll command on the server side, embodiments of the present invention use a Firefox extension to capture all the mouse events and interpret them as certain gestures appropriately.

[0167] According to one embodiment of the invention, the audio decoder 232 module decodes the audio streamed from the server. In one embodiment of the invention the audio decoder 232 will use the native audio decoder of the mobile platform. In another embodiment of the invention the audio decoder 232 will use the CELT decoder library. The CELT codec is a low latency compression-decompression algorithm for audio.

[0168] In one embodiment of the present invention, the player 236 and the other modules the player 236 uses on the client device 124, including the video decoder 230 and audio decoder 232, are packaged as a native application for a particular client device. For example, there is a player application written specifically for iOS devices, such as iPhone 126 or iPad 128 mobile devices, and another player application written specifically for Android 130 mobile devices.

[0169] FIGS. 3A-3B are flow diagrams of a process for providing interactive content generated by an application configured to run in one application environment, to a client device providing a different application environment, according to one embodiment of the invention.

[0170] The process starts, and the monitoring server 114 receives from the client device 124 identification of the particular software application to be invoked. In this regard, the monitoring server 114 receives a user selection of the particular application in step 310.

[0171] In step 312, the player 236 on the client device 124 sends a request for the selected application to the configurator 214 on the monitoring server 114.

[0172] In step 314 the configurator 214 returns the appropriate configuration file for the selected application back to the player 236, and the player 236 configures itself based on the parameters of the returned configuration file. For example, the player sets its display resolution, maximum bit rate, maximum frame rate, audio and video codec used, and the like.

[0173] In step 316, the player 236 sends a request to the load balancer 216 on the monitoring server 114 to start an instance of the selected application.

[0174] In step 318, the load balancer identifies and selects a specific delivery server 112 and an instance of the provisioner 218 on the same delivery server. According to one embodiment, the load balancer may select a delivery server which is geographically close to the client device.

[0175] In step 320 the provisioner 218 responds to the load balancer 216, who responds to the player 236, with parameters that include credentials for accessing the created instance of the delivery server 112, an IP address of the

delivery server, and an audio port on the delivery server, to which the player should connect.

[0176] In step 322, the player 236 uses the parameters returned in step 320 to connect on the control channel to the instance of the control server 220 selected for the instance of the application.

[0177] In step 324 the control server 220 and the player 236 exchange over the control channel, parameters (e.g. username and password) for the instance of the application, frame rate, bit rate, supported video and audio codec, and the like.

[0178] In step 326 the control server 220 responds to the player 236 over the control channel with the video port on the delivery server 112 in which the player should connect.

[0179] In step 328 the player 236 connects to the video port specified by the control server 220 and the audio port specified by the provisioner 218.

[0180] In step 330 the control server 220 invokes the application in the application environment provided by the delivery server 112 and notifies the player 236 that the selected application has finished launching.

[0181] In step 332 the selected application generates screen display outputs and provides the outputs on the virtual screen 224 on the delivery server 112.

[0182] In step 410 the video capture and streamer 226 receives video parameters, such as, for example, frame rate, bandwidth, bit rate, and video port, from the control server 220.

[0183] In step 412 the video capture and streamer 226 captures and encodes a screen display output by the virtual screen 224 into various video frames according to the received frame rate. According to one embodiment, in order to achieve minimal latency in the system, each video frame is encoded after being captured, and the output from the encoder is fed to the video port without excessive buffering or delay. According to one embodiment, the maximum frame size of each video frame is based on the specified frame rate and the maximum bit rate.

[0184] In step 414 the audio capture and streamer captures audio from the audio sink of the selected application and then encodes and streams the audio to the client device 124 over the audio port.

[0185] In step 416 the video decoder 230 and audio decoder on the client device 124 respectively decodes the received video stream and audio stream, which the player 236 then displays or plays respectively on the client device. In this regard, the video decoder continuously parses the incoming video data from the socket. As soon as it collects all the data required to decode a video frame, the frame is decoded, resized if needed, converted to an appropriate color space, and displayed. Again, video data are decoded and displayed as soon as they are available and there is no excessive buffering to ensure minimal latency. According to one embodiment, the video decoder starts the decoding as soon as it identifies a tag (e.g. an AUD tag) which is inserted at the end of a particular video frame.

[0186] In step 418, the user enters input to the selected application via the player 236 using touches, gestures, keyboard entry, or any other form of input. According to one embodiment, the player 236 maps the user input (e.g. downward swipe, shaking of the device, changing orientation of the device, tapping on the screen, and the like) to a particular command defined in the configuration file for the application provided to the client device. For example, changing the orientation of the device may be mapped to a “sendOrienta-

tionChanged” command which receives, as a parameter, the orientation of the device. The change of orientation of the device and other inputs to the device are interpreted according to conventional mechanisms that will be understood by a person of skill in the art.

[0187] In step 420, the player 236 sends the mapped command(s) over the control channel to the control server 220.

[0188] In step 422 the control server converts the commands received from the player 236 to a corresponding input catered to the particular application (e.g. mouse or keyboard command understood by the application), and forwards the converted input commands to the application in the application environment 222.

[0189] In step 424, the selected application reacts to the user input or internal events, and changes its screen and sound output appropriately. In step 426 the changes to the screen are reflected onto the virtual screen 224 and the changes to the sound to the audio sink. The sequence then goes to step 412, and the streaming of video and audio to the client device continues.

[0190] In one embodiment of the invention, instead of mapping the user input to predefined high-level gestures or commands, such as sending a scroll down command when the user swipes down on the screen, the simplest mapping of the input or even the raw touch/input data from the client device is sent to the delivery server 112 over the control channel. For example, a mouse down event is sent when a user touches down on the screen. This mapping of input to the simplest command expected by the application environment on the server side avoids mapping to more complicated actions, such as high-level gestures or commands. The raw user inputs are then interpreted differently based on the content inside the application environment 222 and/or the context of the application. For example, a Firefox application provides a different context, which is browser based, than an Android application, which is touch-based. Thus, a mouse-down event may be transmitted to a Firefox application when a same action by a user would create the transmitting of a touch down event for an Android application. With respect to the transmitting of input based on content, instead of interpreting what the user wants to do (via, for example, gesture recognizers) on the client device, and sending it to the application on the server side, according to one embodiment of the invention, such gestures and/or user intentions are interpreted on the server side. For example, a touch down immediately followed by a series of touch moves and a touch up (i.e. a swipe/drag action) on the mobile device would imply that the user wants to scroll the page. If the control server makes a minimal interpretation based on the context of the application and transmits a mouse down input followed by mouse moves and mouse up inputs to, for example, a Firefox application, the application may interpret the input as a selection action instead of a scroll. Thus, for certain applications, the server is configured to capture all the mouse events and interpret them as a gesture prior to transmitting the input to the application. For a Firefox application environment, for example, whether interpreting a touch down event followed by a series of touch move events and a touch up event, as a scroll command, depends on the move direction. In the embodiment of the invention for Flash apps, such interpretation as a scroll command is disabled when the mouse down happens inside the Flash object, where the Flash object itself would interpret these mouse down/move/up events. Thus, for a Flash object, the inputs are not changed,

and the minimally interpreted input (e.g. mouse down->move->up) is sent to the Flash object for interpreting the command on its own.

[0191] Interpreting the touch events on the delivery server 112 has the advantage of achieving content-aware gesture mapping. The application environment 222 knows exactly where the user touches and thus where the gesture starts and ends. Therefore the gesture may be better mapped creating a better user experience. In addition, sending the raw touch data improves responsiveness of the application because the raw input data is provided as the inputs are being entered instead of waiting for the entire input to complete. For example, there is no need for a swipe down event to complete before transmitting such an event to the application.

[0192] In one embodiment of the invention, the control channel is configured to transmit multiple touch inputs for application environments 222 that accept and expect such events (e.g. an Android application running in the delivery server 112). In this case, a touch tracker in the player 236 tracks the multiple touches. Each touch at the client side is assigned with a touch ID. According to one embodiment, each finger is represented with a touchID having associated coordinates (x, y). Each time the user starts or ends a touch event (e.g. putting one more finger down or lifting one finger up), the touch tracker in the player 236 groups all the touches along with the action (down/move/up) and the active touch ID that corresponds to the finger that initiated the event. This data is then sent through the control channel and to the control server 220. The control server 220 organizes the touch data into the format accepted by the application environment 222, and sends the formatted data to the application environment 222.

[0193] FIG. 4 a flow diagram of process for dynamically adjusting the quality of streamed video frames during a streamed video session based on network congestion over a TCP socket, according to one embodiment of the invention. The process implements the QoS algorithm of table I. According to one embodiment, the process is implemented by the video capture and streamer module 226 on the mobile app delivery server 112.

[0194] In one embodiment of the invention, video quality is either dynamically downgraded or upgraded based on the concurrent congestion status of a TCP socket. In this regard, the process adjusts video quality by varying two aspects of video quality: video frame rate and video bit rate. For example, the video frame rate may range from 10 to 24 frames per second (fps), and the video bit rate may range from 320K to 2048K bits per second (bps). Video frame rate affects primarily the smoothness or choppiness of a video. For a given video frame rate, the video bit rate affects primarily the visual quality of individual frames.

[0195] In one embodiment of the invention, as congestion over a TCP socket occurs, both the video frame rate and video bit rate are downgraded together until the video bit rate hits a predefined floor value MIN_FR, after which the video frame rate is held constant but the video bit rate continues to be downgraded.

[0196] According to one embodiment, the process of FIG. 4 is invoked before the encoding of each and every frame of video to be transmitted to the client device. However at the initialization of the video streaming session, certain constant values are set that control the range and frequency of the adjustment of video quality. In one embodiment of the invention, the following constants are used:

[0197] MIN_LEVEL=minimum QoS level (e.g. 5);

[0198] MAX_LEVEL=maximum QoS level (e.g. 32);

[0199] MIN_FR=minimum frame rate (e.g. 10);

[0200] MAX_FR=for maximum frame rate (e.g. 24);

[0201] MIN_MBR=for minimum bit rate (e.g. 320K);

[0202] MAX_MBR=maximum bit rate (e.g. 2048K);

[0203] MBR_STEP=bit rate increment (e.g. 64K);

[0204] DOWNGRADE_INTERVAL=downgrade interval time (e.g. 1000 milliseconds); and

[0205] UPGRADE_INTERVAL=upgrade interval time (e.g. 1000 milliseconds).

[0206] The outcome of the process is the setting of the video frame rate (fr) and the video bit rate (mbr) parameters which are subsequently be used by the video encoder to encode the next video frame.

[0207] The process in FIG. 4 starts, and in step 510, the video capture and streamer module 226 before the encoding of each frame first sets a previous quality level to the value of a current quality level, and further sets a current time to the current system time.

[0208] In step 512, the video capture and streamer module 226 determines whether the TCP socket was congested after the transmission of the last video frame. If the TCP socket was congested, the process goes to step 514 where the process attempts to downgrade the quality level by adjusting the value of the current quality level. In this regard, the process sets the value of a last congestion time to the value of the current time, and increments a congestion counter for the current quality level.

[0209] In step 516, the process determines whether the current quality level is greater than a preset minimum quality level. If the answer is NO, the process proceeds to step 524 because the process does not downgrade the level below the preset minimum quality level.

[0210] If the current quality level is greater than the preset minimum quality level, the process determines whether the difference between the current time and the last downgrade time exceeds a preset downgrade interval. If it does not, the processor proceeds to step 524. According to one embodiment, using a downgrade interval to control how frequently the level is downgraded insures a smoother downgrade process without wild down swings. If the downgrade interval time has been exceeded, the process proceeds to step 518.

[0211] In step 518, the video capture and streamer module 226 downgrades the current quality level to the preset minimum quality level if the current quality level is less than or equal to a preset minimum frame rate. If the current quality level is greater than the preset minimum frame rate, the process downgrades the current quality level to either the preset minimum frame rate or the preset minimum QoS level, whichever is a greater value. In either case, after the downgrade has been done, the process proceeds to step 524.

[0212] Referring again to step 512, if a determination is made that the TCP socket was not congested, the video capture and streamer module 226 proceeds to step 520 to check if an upgrade of the video quality level is possible. In this regard, the process determines if the current quality level is less than a preset maximum QoS level. If the answer is NO, meaning that the current quality level is already at the maximum level, the process proceeds to step 524.

[0213] If the current quality level is less than the preset maximum QoS level, the process proceeds to step 522 to attempt to upgrade the level by a preset amount, such as, for example, by one. In this regard, the process uses a congestion

counter for the current quality level to compute a multiplier. According to one embodiment, if the congestion counter is zero, the multiplier is one; if the congestion counter is one, the multiplier is two; if the congestion counter is two, the multiplier is four, and so on. According to one embodiment, the multiplier acts as an exponential dampener. The multiplier is multiplied by a preset upgrade interval time to compute an upgrade delay time. The current time is then subtracted from the greater of either the last upgrade time or the last congestion time, and if the difference has exceeded the computed upgrade delay time, an upgrade take place. If an upgrade is allowed, the current quality level is incremented by one. In either case, the algorithm then proceeds to step 524.

[0214] In step 524, the video capture and streamer module 226 checks if either a downgrade or an upgrade of the current quality level has just occurred. If not, the process proceeds to step 532. If an upgrade or downgrade has just occurred, the process determines, in step 526, whether a downgrade has occurred. If the answer is YES, the process proceeds to step 528 where the last downgrade time is set to the current time and the process proceeds to step 532.

[0215] If instead in step 526 it is determined that an upgrade has just occurred, the process proceeds to step 530 for setting the congestion counters to zero for all quality levels less than or equal to the current level. That is, because of the upgrade, the process allowed to have confidence in the lack of congestion in all levels less than or equal to the current level to which it has been upgraded. The algorithm then proceeds to step 532.

[0216] Step 532 is the culmination of the process where the video frame rate and video bit rate are set based on the results of the algorithm. According to one embodiment, the video bit rate is set to the current quality level multiplied by a preset bit rate increment. In computing the video frame rate, the maximum value of either the current quality level or the preset minimum frame rate is selected. The result is then compared against the preset maximum frame rate. The minimum of the two values is then set as the video frame rate. The process ends, and the video encoder and streamer module 226 uses the set video frame rate and video bit rate to encode the next video frame to be streamed to the client device.

[0217] The processes of FIGS. 3-4 may be described in terms of a software routine executed by the corresponding CPU based on instructions stored in memory. A person of skill in the art should recognize, however, that the processes may be executed via hardware, firmware (e.g. via an ASIC), or in any combination of software, firmware, and/or hardware. Furthermore, the sequence of steps of the processes are not fixed, but may be altered into any desired sequence as recognized by a person of skill in the art.

[0218] A person of skill in the art should appreciate the embodiments of the present invention allow the efficient delivery of interactive applications built for one environment to various mobile devices (and other clients) for which they were not originally written. The various embodiments allow the applications to function on the new platforms with little or no redevelopment of the applications. Application developers will be able to quickly re-purpose their existing application assets to run on the new environments without the need to learn the technical details of the new platforms. The above embodiments also help eliminate the need to download the applications to the new platforms, or purchase separate applications for the new platforms.

[0219] According to the above embodiments, the applications are configured to appear to function on the new platform in the same way as they would on the native platform. The embodiments of the present invention also allow the applications to be highly responsive to the user's control. The high responsiveness is aided by reducing the latency of upstream inputs from the client to the server running the application, and latency of downstream audio and video from the server to the client.

II. Interactive App Discovery

[0220] Embodiments of the present invention provide a system and method for experiencing advertised interactive applications. Such apps may be, for example, interactive gaming applications, and the clients may be, for example, mobile devices 124 (see FIG. 1). However, a person skilled in the art should recognize that any app may be so discovered and other platforms besides mobile devices are covered by embodiments of the present invention, including but not limited to the Web, TV platforms, and the like.

[0221] In general terms, when the client is a mobile device, a method for discovering, advertising, experiencing, demoing and purchasing an app using the present invention includes:

[0222] 1. On a mobile device, the end user starts a first interactive application (publisher app), which is resident on and native to the mobile device, and which may provide some function like a game, as well as hosting second interactive applications (AppGlimpse ad units) by including an advertising player module (AppGlimpse SDK Library) in its executable.

[0223] 2. The publisher app's game plays and at the completion of some game level (or some other trigger point), the publisher app looks for and detects a condition and determines that an AppGlimpse ad unit is to be played for the user.

[0224] 3. The publisher app calls the AppGlimpse SDK library, which is linked with its executable on the mobile device, and requests that an AppGlimpse ad unit be played.

[0225] 4. The controller component of the AppGlimpse SDK library performs initialization actions including contacting mobile app delivery system servers to select an app server to virtually play a selected AppGlimpse ad unit for an advertised app.

[0226] 5. The selected AppGlimpse ad unit for the selected advertised app begins execution in a virtualization environment of the mobile app delivery system servers. The presentation of the app is streamed as video, audio, and/or OpenGL frames/commands, to the AppGlimpse player in the SDK library of the publisher app on the mobile device.

[0227] 6. The AppGlimpse player on the client mobile device decodes the streamed information and presents the AppGlimpse ad unit to the user. It also captures and forwards all interactive input (interactive data) by the user and sends that information over a control channel to the app servers in the mobile app delivery system.

[0228] 7. Using the app virtualization environment on the app servers, the inputs are fed into the AppGlimpse ad unit, as if the advertised app were running in its native computing environment (including native operating system), and the AppGlimpse Ad unit responds to the user's input and displays the next frames.

[0229] 8. This back and forth between client and server continues until the AppGlimpse ad unit session is complete, at which time control is returned to the publisher app on the client mobile device.

[0230] FIG. 5 is an overview block diagram of components of a system for experiencing advertised interactive apps, and messages exchanged by those components, according to one embodiment of the present invention

[0231] According to one embodiment of the invention, the system includes a mobile device computing environment 710 in communication with a mobile app delivery system environment 711 over a data communications network such as, for example, a local area network, private wide area network, or a public wide area network such as, for example, the Internet. The mobile app delivery system environment may be a computing environment similar to the mobile app delivery system environment 110 described above with respect to FIG. 2. The mobile app delivery system environment 711 includes, but is not limited to, a set of load balancer servers 722, a set of services servers 724, a set of app servers 726, and a set of ad units 728. The app server 726 may be similar to the mobile app delivery servers 112 which are described above with respect to FIG. 2, and includes but is not limited to all the components contained within the mobile app delivery servers 112 which are also described above with respect to FIG. 2. The app servers 726 also correspond to the virtualization and streaming servers referred to earlier above. The virtualization environment may be similar to the app environment 222 previously described with respect to FIG. 2, along with the Virtual Screen 224, Video Capture and Streamer 226, Audio Capture and Streamer 228 components that support it also previously described above with respect to FIG. 2. The load balancing servers 722 may be similar to the load balancer 216 included within the mobile app monitoring server 114, described above with respect to FIG. 2.

[0232] According to one embodiment, the app server is configured to select an ad for an interactive application in response to a request from the mobile device environment. According to one embodiment, the ad units 728 are interactive advertised apps or portions of interactive advertised apps packaged as AppGlimpse ad units and stored in a storage device.

[0233] According to one embodiment, the mobile device environment 710 may be a computing environment hosting a set of publisher app modules 716 containing executable code for one or more publisher apps, and executable code for the advertising player module (AppGlimpse SDK library) 714. The AppGlimpse SDK library 714 in turn includes but is not limited to a controller component 718 and a player component 720. According to one embodiment, the controller and player components may be deemed to be embedded in the publisher app module 716. According to one embodiment, the publisher app 716, controller component 718, and player component 720 are implemented via computer program instructions which are stored in memory and executed by a processor of the mobile device.

[0234] According to one embodiment of the invention, within the mobile device environment 710, a particular one of the one or more publisher apps 716 is selected by a user by invoking an input unit of the mobile device, and the selected publisher app begins to execute. The app may be, for example, an interactive game app. At some point in its execution, the publisher app 716 detects a condition for invoking an ad unit, and, in act 752, executes an init call to the controller component 718 of the AppGlimpse SDK Library 714 for invoking the ad unit. The condition may be for example, lapse of certain amount of time in interacting with the publisher app, reaching

a certain preset milestone in the publisher app, detecting a user command to end the publisher app, and/or the like.

[0235] In act 754, the controller 718 transmits an init and configure request over the data communications network to the services server 724 in the mobile app delivery system environment 711.

[0236] In act 756, the services server 724 returns a response to the controller 718, which includes, but is not limited to, a list of load balancer servers 722 and configuration parameters. The configuration parameters include but are not limited to frame rates, bit rates, quality of service data, and the like.

[0237] Still referring to FIG. 5, in one embodiment of the invention, the controller 718, in act 758, sends a series of ping requests to the list of load balancers 722 and receives a series of ping responses 759. According to one embodiment, the controller 718 selects a specific load balancer 722 providing the quickest ping response for later use.

[0238] In act 760, the controller 718 returns control to the publisher app 716.

[0239] Still referring to FIG. 5, in one embodiment of the invention, at some point in its execution, the publisher app 716, in act 762, calls the controller 718 to play an ad. The ad may be, for example, for another interactive application such as, for example, a game application.

[0240] In act 764, the controller 718 sends a request over the network to the selected load balancer 722 to play the ad.

[0241] In act 766, the load balancer 722 sends a request to the services server 724 to select an ad. The selection criteria for ads includes but is not limited to: (1) the ad publisher app that is invoking the ad; (2) the type and genre of the publisher app; (3) other apps the user already has installed on their device; (4) filters applied by the publisher (e.g. don't show apps X, Y, Z); (5) filters applied by the advertiser (e.g. don't show my ad on apps X, Y, Z); and/or (6) behavioral information acquired about the user.

[0242] In act 768, the services server 724 returns an ID of a selected AppGlimpse ad unit for an advertised app.

[0243] In act 770 the load balancer 722 sends a provisioning request to a selected app server 726, where the app server is selected based on current load.

[0244] In act 772, the selected app server 726 sends a response containing app server credentials to the load balancer 722. The app server credentials may include information such as IP address, port, and the like.

[0245] In act 774 the load balancer 722 sends a response to the controller 718, which includes the app server credentials and ad ID information.

[0246] In act 776, the controller 718 initiates a background action of downloading the closing frame by sending a request to the services server 724.

[0247] In act 778, the services server 724 returns a closing frame response. The closing frame may be a video frame, a native frame written in the native language of the client mobile device, an HTML page, or any other data for displaying an interactive screen to the user. The closing frame includes information on what to display to the user after the AppGlimpse ad session is over such as, for example, prompts to "Buy", "Replay", "Close," and the like, and the actions to be taken when a choice is made. The selection criteria for the closing frame includes but is not limited to: (1) the advertised app; (2) the publisher app; (3) the call to actions we want to include (e.g. buy, replay, close etc.); and (4) the particular section of the app that was played in the ad unit for the

advertised app. According to one embodiment of the invention, the closing frame is downloaded independently of the frame from the ad unit, and at any point in time. According to one embodiment, display of the closing frame is invoked upon receipt of a final frame of the ad unit.

[0248] Still referring to FIG. 5, in one embodiment of the invention, in act 780, the player 720, using the app server credentials and ad ID, sends a request to the selected app server 726 to start the AppGlimpse ad unit for the selected advertised app. According to one embodiment, prior to start play of the ad unit, a prompt may be displayed to the user asking if the user is interested in interacting with the advertised app.

[0249] Still referring to act 780 a back and forth dialog ensues between the player 720 on the client device and the app server 726 to play and stream the advertised app, as is described in further detail in section I above. According to one embodiment, the app server 726 executes the selected ad unit 728 in its app virtualization environment and then streams video, audio, or OpenGL frames or commands to the player 720 on the client. According to one embodiment, the app server 726 generates a display output that is captured into video frames and streamed to the player component 720. In another embodiment, OpenGL rendering commands are generated and streamed to the player component. The player 720 decodes the video, audio, or OpenGL frames or commands and presents the decoded data to the user (e.g. via a display screen). The player 720 also collects all input, including user interaction data, and forwards the input to the app/game server 726 over a control channel over the network. In addition, as outputs generated by the ad unit 728 is streamed to the client at various points in act 784, the output is reported to the services server 724 for tracking progress of the app.

[0250] In act 786, the app server 726 sends a shutdown command over the network to the player 720, which returns control to the controller 718. According to one embodiment, the shutdown command may be in response to the AppGlimpse ad unit session being completed. The session may be deemed to be complete when the advertised application finishes executing, or in response to a specific user command received prior to the advertised application finishing its execution, such as, for example, a command to end the ad.

[0251] In act 790, the controller 718 shows the closing frame 732 on a display of the mobile device in response to the shutdown command. The closing frame may include, for example, various interactive options relating to the advertised app, such as, for example, options for the user to buy, replay, or close the advertised app, and/or provide feedback or rating of the advertised app. In some embodiments, options unrelated to the advertised app may also be displayed in the closing frame. The user provides input indicative of a selection of one of the displayed options, and the controller component 718 is configured to process the user input and generate an output. According to one embodiment, the output may be connecting the mobile device to a server, such as, for example, an app store server 730. For example, if the user selects a buy option in act 792, in one embodiment of the invention, the user is redirected to the appropriate app store server 730 to begin the purchase and download process for the advertised app.

[0252] In act 798, at the end of processing for the closing frame 732, the controller 718 returns control to the publisher app and the user continues to interact with the publisher app as he or she did prior to invoking the ad. For example, the user

may resume play of the game in the case that the publisher app is a gaming app. As a person of skill in the art should appreciate, the user gets to enjoy and interact with the advertised app without having to leave the current app. For example, the user may play the advertised app during the middle of a current game provided by the current app, and seamlessly return to the current game after the advertised app has been sampled to the user's satisfaction. The sampling of the advertised app is possible even if the advertised app is configured to run in a different computing environment, including, for example, in different operating systems.

[0253] Each of the various servers, applications, and/or modules in the afore-described figures may be a process or thread, running on one or more processors, in one or more computing devices, executing computer program instructions and interacting with other system components for performing the various functionalities described herein. The computer program instructions are stored in a memory which may be implemented in a computing device using a standard memory device, such as, for example, a random access memory (RAM). The computer program instructions may also be stored in other non-transitory computer readable media such as, for example, a CD-ROM, flash drive, or the like. Also, a person of skill in the art should recognize that the instructions may be implemented via firmware (e.g. an application-specific integrated circuit), hardware, or a combination of software, firmware, and hardware. A person of skill in the art should also recognize that the functionality of the servers, applications, and/or modules may be combined or integrated into a single server, application, and/or module, or distributed among various servers, applications, and/or modules in a single or distributed computing environment without departing from the scope of the exemplary embodiments of the present invention.

[0254] According to one embodiment, the various servers are hosted using infrastructure in a remote cloud-based computing environment. Also, the various servers may run as virtual machines implemented as a software-based computer running on a physical machine.

[0255] Although this invention has been described in certain specific embodiments, those skilled in the art will have no difficulty devising variations to the described embodiment which in no way depart from the scope and spirit of the present invention. Furthermore, to those skilled in the various arts, the invention itself herein will suggest solutions to other tasks and adaptations for other applications. For example, although embodiments of the present invention contemplate providing content from applications configured to run on a particular application environment, to mobile platforms providing a different application environment, a person of skill in the art should recognize that the sender and recipient devices may run the same application environment. Also, the receiving platform need not be a mobile platform but any computing platform known in the art.

[0256] Also although embodiments of the present invention contemplate that the advertisements will be for interactive applications, a person of skill in the art should recognize that the advertisement may be for movies, music, or other interactive and non-interactive applications conventional in the art.

[0257] It is thus the applicant's intention to cover by claims all such uses of the invention and those changes and modifications which could be made to the embodiments of the invention herein chosen for the purpose of disclosure without

departing from the spirit and scope of the invention. Thus, the present embodiments of the invention should be considered in all respects as illustrative and not restrictive, the scope of the invention to be indicated by the appended claims and their equivalents rather than the foregoing description.

1. A method for experiencing advertised interactive applications, comprising:

invoking, by a processor, a first interactive application, wherein the first interactive application includes an embedded advertising player module;

receiving, by the processor, first interaction data for interacting with the first interactive application;

invoking, by the processor, the advertising player module in response to the first interactive application detecting a condition;

transmitting, by the processor via the advertising player module, a request for a second interactive application, wherein in response to the request, the second interactive application is invoked by a first server coupled to the processor, the second interactive application generating a screen display output, wherein in response to the screen display output, video frames or rendering commands are streamed to the advertising module;

receiving, by the processor via the advertising player module, the streamed video frames or rendering commands and generating a display in response;

receiving, by the processor via the advertising player module, second interaction data for interacting with the second interactive application; and

forwarding, by the processor via the advertising player module, the second interaction data to the first server, wherein the second interactive application provides an output in response to the second interaction data.

2. The method of claim 1 further comprising:

displaying, by the advertising player module, an option relating to the second interactive application;

receiving, by the advertising player module, a user input indicative of a selection of the option; and

connecting, by the advertising player module, in response to the user input, to a second server configured to generate an output responsive to the selected option.

3. The method of claim 2, wherein the option is to purchase the second interactive application.

4. The method of claim 2 further comprising:

receiving by the advertising player module a message indicative of an end of the second interactive application;

in response to receipt of the message, displaying by the advertising player module the second video frame including the option relating to the second interactive application.

5. The method of claim 4 further comprising:

detecting an end of processing relating to the selected option; and

invoking, by the processor, the first interactive application for resuming interaction with the first interactive application.

6. The method of claim 1, wherein the native computing environment of the first interactive application is different from the native computing environment of the second interactive application.

7. The method of claim 6, wherein each native computing environment provides an operating system, and the operating system in which the first interactive application is configured

to run is different than the operating system in which the second interactive application is configured to run.

8. The method of claim 1, wherein the processor receives the streamed video frames or rendering commands over a wide area network.

9. An apparatus for experiencing advertised interactive applications, the apparatus comprising:

an input interface for receiving input commands;

an output interface for providing audio or visual output;

a processor; and

memory, wherein the memory stores instructions that, when executed by the processor, cause the processor to:

invoke a first interactive application, wherein the first interactive application includes an embedded advertising player module;

receive first interaction data via the input interface for interacting with the first interactive application;

invoke the advertising player module in response to the first interactive application detecting a condition;

transmit, via the advertising player module, a request for a second interactive application, wherein in response to the request, the second interactive application is invoked by a first server coupled to the processor, the second interactive application generating a screen display output, wherein in response to the screen display output, video frames or rendering commands are streamed to the advertising module;

receive, via the advertising player module, the streamed video frames or rendering commands and generating a display via the output interface in response;

receive, via the advertising player module, second interaction data for interacting with the second interactive application; and

forward, via the advertising player module, the second interaction data to the first server, wherein the second interactive application provides an output in response to the second interaction data.

10. The apparatus of claim 9, wherein the instructions further cause the processor to:

display, via the advertising player module, an option relating to the second interactive application;

receive, via the advertising player module, a user input indicative of a selection of the option; and

connect, via the advertising player module, in response to the user input, to a second server configured to generate an output responsive to the selected option.

11. The apparatus of claim 10, wherein the option is to purchase the second interactive application.

12. The apparatus of claim 10, wherein the instructions further cause the processor to:

receive, via the advertising player module, a message indicative of an end of the second interactive application;

in response to receipt of the message, display, via the advertising player module, the second video frame including the option relating to the second interactive application.

13. The apparatus of claim 12, wherein the instructions further cause the processor to:

detect an end of processing relating to the selected option; and

invoke the first interactive application for resuming interaction with the first interactive application.

14. The apparatus of claim **9**, wherein the native computing environment of the first interactive application is different from the native computing environment of the second interactive application.

15. The apparatus of claim **14**, wherein each native computing environment provides an operating system, and the operating system in which the first interactive application is configured to run is different than the operating system in which the second interaction application is configured to run.

16. The apparatus of claim **9**, wherein the processor receives the streamed video frames or rendering commands over a wide area network.

* * * * *