University of Wollongong

Research Online

Faculty of Informatics - Papers (Archive)

Faculty of Engineering and Information
Sciences

2007

Distributed Server Delivery of Immersive VoIP Service: Balancing Between Voice Latency and Access Upload Cost

Ying Peng Que

University of Wollongong, ypq01@uow.edu.au

Farzad Safaei

University of Wollongong, farzad@uow.edu.au

P. Boustead

University of Wollongong, boustead@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/infopapers>

Part of the [Physical Sciences and Mathematics Commons](#)

Recommended Citation

Que, Ying Peng; Safaei, Farzad; and Boustead, P.: Distributed Server Delivery of Immersive VoIP Service: Balancing Between Voice Latency and Access Upload Cost 2007.
<https://ro.uow.edu.au/infopapers/629>

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Distributed Server Delivery of Immersive VoIP Service: Balancing Between Voice Latency and Access Upload Cost

Abstract

As a future communication tool to the Distributed Virtual Environment, the Immersive VoIP service delivers to each user a real-time mix of nearby user voices which are all rendered according to their respective virtual world positions. In this work, we conjecture that the voice latency constraints between Immersive VoIP users would vary according to their virtual world distance of separation. As verified in our subjective listening tests, this conjecture is more realistic than fixed latency constraint, in terms of user perceptual experience and network feasibility. Empowered by the Bandwidth Reduced server assignment algorithm, the proposed two-hop distributed server architecture achieves a better balance between access bandwidth cost and voice latency when compared against the central server and the peer-to-peer architectures. The two-hop architecture targets the challenging scenario where there lacks clear correlations between the virtual world and the underlying physical network. In addition, we have also investigated the necessary frequency to execute the server assignment solutions to mitigate sensitivity to virtual world mobility.

Disciplines

Physical Sciences and Mathematics

Publication Details

This conference paper was originally published as Que, YP, Safaei, Z, Boustead, P, Distributed Server Delivery of Immersive VoIP Service: Balancing Between Voice Latency and Access Upload Cost, 15th IEEE International Conference on Networks ICON 2007, 19-21 Nov, 489-494.

Distributed Server Delivery of Immersive VoIP Service: Balancing Between Voice Latency and Access Upload Cost

Que Y. P., Safaei F. and Boustead P.

Smart Internet CRC,

Telecommunications and Information Technology Research Institute,

University of Wollongong, Australia

Email {ying, farzad, paul}@titr.uow.edu.au

Abstract—As a future communication tool to the Distributed Virtual Environment, the Immersive VoIP service delivers to each user a real-time mix of nearby user voices which are all rendered according to their respective virtual world positions. In this work, we conjecture that the voice latency constraints between Immersive VoIP users would vary according to their virtual world distance of separation. As verified in our subjective listening tests, this conjecture is more realistic than fixed latency constraint, in terms of user perceptual experience and network feasibility. Empowered by the Bandwidth Reduced server assignment algorithm, the proposed two-hop distributed server architecture achieves a better balance between access bandwidth cost and voice latency when compared against the central server and the peer-to-peer architectures. The two-hop architecture targets the challenging scenario where there lacks clear correlations between the virtual world and the underlying physical network. In addition, we have also investigated the necessary frequency to execute the server assignment solutions to mitigate sensitivity to virtual world mobility.

I. INTRODUCTION

Recently, there has been an increasing trend in applying Voice over IP (VoIP) service to support group communications in Distributed Virtual Environment (DVE). This class of applications is epitomised by the voice chat services integrated with the network game engine Xbox Live [1]. In a DVE such as that of a Massively Multi-player Online Games (MMOG) [2], each user is represented by an *avatar* in the virtual world. In our work, we assume a one to one correspondence between an avatar in the virtual world and a particular client (wired or wireless) in the underlying physical network. As shown in Fig. 1a, a given avatar has a *communication zone* of which the radius is defined by the avatar's hearing range. We define communication density as the average number of neighbouring avatars residing in each communication zone. One of the main attractions of DVE is the ability to enhance the users' social interactive experience through creating a sense of virtual world immersion [3]. As shown in [4], the current generation of single channel (mono) VoIP services such as Xbox LIVE, seems disconnected from the visual DVE scenes and can not contribute to the users' sense of immersion. As a future development, we propose to deliver the Immersive VoIP service [5] to complement the visual DVE scenes and further enhance the users' sense of immersion. In Immersive VoIP service, a personalised

Auditory Scene is created for each avatar which is the real-time mixture of all the voices heard by that avatar as defined by its *communication zone*. In an *Auditory Scene*, each voice stream is localised (directional placement) and distance-attenuated with respect to the appropriate virtual world avatar positions, thus creating a virtual "*cocktail-party*" effect [6].

The Immersive VoIP service often has to cater for the concurrent access of a large population of densely populated avatars. More importantly, avatars in close virtual world proximity can be controlled by clients located far apart in the physical network. It is therefore important to deliver the Immersive VoIP service with a balance between scalability and quality. Three voice quality requirements have been identified in [7], i.e., directional voice rendering accuracy, voice attenuation accuracy and voice communication latency. In [5], the scalability of Immersive VoIP service has been significantly improved in terms of access bandwidth and voice processing costs while still maintaining satisfactory level of the first two quality requirements. In this work, we focus on reducing voice latency while still maintaining reasonable level of scalability. To accurately measure voice latency, we conjecture that the delay constraints between different Immersive VoIP users are governed by their virtual world distance of separation. The perceptual realism of this conjecture is verified in our subjective listening tests (Fig. 2).

In the full-mesh Peer-to-peer (P2P) architecture, peers directly exchange voice between each other. The full-mesh and other improved types of P2P architectures [7] have been commonly adopted to deliver the current VoIP solutions [15]. On the other hand, many existing DVEs, especially network game services employ a central server to manage the state information [15]. Full-mesh P2P achieves the lowest network-feasible (determined by node-to-node delays) bound on voice latency, but at the expense of incurring the maximum bound on access upload and download cost. On the contrary, central server sets the lowest bound on access upload cost (minimum of one stream per user) and access download cost (determined by the playback mechanism, e.g., 2 for HRTF localisations [5]), but at the expense of incurring the maximum bound on voice latency. In this work, we propose the two-hop distributed server architecture where the *Auditory Scene Creations* are carried out across multiple server sites chosen to be close to the Point Of Presences (POPs). The proposed *Bandwidth Reduced Server Assignment Algorithm* enables the two-hop distributed server architecture to achieve a better balance between voice latency and access bandwidth cost. For

instance, at communication density of 30, the *Bandwidth Reduced Algorithm* can reduce the percentage of avatar pairs with violated delay constraints from 54.44% of central server to 38.99% as well as decrease the average uploads from 26.81 of full-mesh P2P to 3.41 streams per avatar. Moreover, the simulation results prove that the *Bandwidth Reduced Algorithm* is scalable enough to support avatar mobility in the virtual world.

Figure 1. Model of the virtual world applied.

The rest of the paper is organised as follows: Section II motivates this work by stating its application scenario which was not addressed by the related prior art. Section III verifies the conjecture of distance-governed variable delay constraints. Section IV presents an overview of the proposed distributed server architecture and the applied virtual and physical world models. Two server assignment algorithms are derived in Section V. We analyse our server assignment solutions in Section VI before final conclusion in section VII.

II. RELATED WORK AND MOTIVATION

Nguyen *et al.* developed the distributed locale server architecture in [10] for delivering Immersive VoIP to Massive Multiplayer Online Games. In Nguyen's architecture, each avatar uploads its voice stream once to a chosen *Auditory Scene Creation* (ASC) server. When communicating with avatars assigned to the same ASC server, voice transmission takes just two hops, i.e. the 1st hop between the POP of speaking avatar and the common ASC server and the 2nd hop between the ASC server and the POP of listening avatar. This minimum two-hop transmission path is the same as our two-hop DISTRIBUTED server architecture. However, when communicating with avatars assigned to different ASC servers, voice transmission incurs additional delay due to the 3rd hop of server to server forwarding. A large amount of server-to-server voice exchanges occur when there are no clear correlations between the avatar positions in the virtual world and the client network locations in the physical world.

Another important attribute of this uncorrelated DVE scenario is a wide geographical spread of POP locations. A typical and popular uncorrelated DVE is a global online game such as World of Warcraft where English-speaking gamers from Australia and USA can engage in close interactions in the virtual world, despite their geographical distance. Nguyen's analytic results show that his architecture achieves almost no delay improvement over central server in an uncorrelated DVE (correlation factor of zero in [10]). In contrast, our architecture aims to achieve significant delay improvement over central server even in the challenging scenario of uncorrelated DVE. As illustrated in Fig. 3, our architecture eliminates the 3rd hop by always uploading voice streams directly from the POP of speaking avatar to the ASC server of the listening avatar.

An important improvement made by our work over Nguyen's work, is that the *Lowest Feasible Delay* constraint (2) enforces our server assignment solutions to meet the delay constraint of each avatar pair to the best ability of the underlying network. In contrast, Nguyen's objective of server assignment was to minimise the sum of average interactive delays for all avatars. Average interactive delay is defined as the average delay between each avatar and the members of its interactive zone of which the radius is set to a fraction e.g., one third, of hearing range. Consequently, Nguyen's absolute delay based objective does not take into account the variation in delay constraints between different avatars as captured by our concept of distance-governed variable delay constraints (Fig. 1b and Fig. 2). In one particular simulation scenario, we set the transmission latency constraints between 45 and 120 ms and compared the two objectives over 2000 cluster distributed avatars at communication density of 30. Our server assignment objective reduced the percentage of avatar pairs with above zero delay deviations by 10.62% (Metric 1 delay) and reduced the mean delay deviation (Metric 2 delay) by 5.77%, in comparison to Nguyen's solution.

Finally, Nguyen's work did not investigate of the sensitivity of latency driven distributed server assignment solutions to avatar mobility which is addressed in this work (D of Section VI).

III. THE CONJECTURE OF DISTANCE-GOVERNED VARIABLE DELAY CONSTRAINTS

We believe the voice latency constraints of DVE users are dependent on several influential factors, including the genre of DVE applications and the delay contributed by various steps of voice processing and transmission. In addition, we conjecture in this work that the latency constraint of each pair of communicating avatars is dependent on their virtual world distance of separation. Irrespective of the actual delay incurred, the quality of voice communication between a particular pair of avatars is only impaired when their delay constraint is exceeded.

In order to verify this conjecture of variable *latency constraints*, we conducted a series of listening tests with 27 subjects. Each subject watched a video clip showing an avatar moving away from the centre of the subject's view. The video clip was dubbed with the avatar's voice clip which was amplitude attenuated in accordance with the avatar visual movement, i.e., 5-10 m (close distance), 10-20 m (medium distance), and 20-30 m (far distance). The subjects evaluated

the linear model (Fig. 1b) of relaxing latency constraints (imposed on the voice clip) in accordance with increasing virtual world distance (portrayed by the video clip) between the listening subject and the speaking/moving avatar. Using the Mean Opinion Score (MOS) [13] scale, the subjects rated the perpetual accuracy or realism of the latency constraint variations in reflecting the changing virtual world distance between the avatar pair presented.

Fig. 2 shows the MOS ratings within 95% confidence intervals on three ranges of delay constraints, chosen on the basis of two earlier empirical values of 150 milliseconds [8] and 96 milliseconds [9]. Fig. 2 suggests that when compared against fixed single delay constraint, this conjecture of variable delay constraints is more realistic in terms of both user perceptions and feasibility to the underlying network. An overly strict single delay constraint, e.g., 50 milliseconds, is unfeasible to the underlying network, even when the minimum-delay full-mesh P2P architecture is applied. An overly strict single delay constraint is also unnecessary for avatar pairs far apart in the virtual world whose delay constraint is more relaxed than the close by avatar pairs. On the other hand, a very relaxed single delay constraint, e.g., 150 milliseconds, is perceptually inaccurate for many avatar pairs, especially those in close interactions.

Figure 2. The MOS ratings of variable delay constraints

For this paper, we define 3 (fair) as the minimum MOS rating required by a Role Playing Game (RPG) game. This threshold is expected to be higher e.g. at 4 (good), for a faster paced First-Person-Shooter (FPS) game scenario. Applying the threshold of 3 to Fig. 2, we found the range of voice communication delay constraints between 75 and 150 milliseconds to offer a good balance between perceptual realism and network feasibility. We identified two types of latency components contributing to the overall voice communication delay evaluated the constraints in Fig. 2. The first type is the processing delays caused by voice localisation compression and packetisation. The second type is transmission delay including the POP-to-server and the client-to-POP (access link) delays. Due to the fixed POP locations (local exchanges) and the Latency Driven Distribution of servers, out of all these delay components, our server assignment solution can only influence the POP-to-server voice transmission delay (referred to as the *Latency Constraint* from now on). Nevertheless, we assume a total value of 30 milliseconds [17] for the delay components

unaddressed by our server assignment solutions. We thus impose a range of *Latency Constraint* between 45 and 120 milliseconds on our server assignment solutions throughout this work.

IV. SYSTEM OVERVIEW

A. Distributed Server Architecture

Fig. 3 illustrates our distributed server architecture in which each avatar has a particular server assigned to create its *Auditory Scenes*. In practise, our server assignment solution usually only establishes ASC servers at a limited number of sites rather than using up the full set of available sites. This forms the basis of our *Bandwidth Reduced* server assignment algorithm presented in B of Section V. For instance, in the scenario of Fig. 3, out of the four available sites S_{1-4} , only the two sites S_{1-2} are selected to establish ASC servers.

Figure 3. The Distributed Server Model

In Fig. 3, the wired client of avatar A_1 communicates with the wireless clients of avatars A_2 and A_3 . The voice transmission path is always two hops between any pair of communication avatars are two hops, e.g., $A_1 \leftrightarrow S_1 \leftrightarrow A_2$. The client of each avatar must upload its mono voice stream to the ASC servers of all the avatars in the communication zone. For example, A_1 needs to upload to S_1 and S_2 which are the respective ASC servers for A_2 and A_3 . As shown in Figure 4, our *Bandwidth Reduced Algorithm* is able to reduce this additional access upload cost to 3.41 uploads per client at the communication density of 30 while achieving significant latency improvements in comparison to central server and Nguyen's architecture (same performance as central server [10] in uncorrelated DVEs). As shown in Fig. 3, our two-hop distributed server architecture has the same minimum access download cost as central server, i.e., each avatar only downloads from one particular ASC server. Therefore only the access upload cost needs to be minimised in Objective Function (1) which is measured as the average number of voice streams uploaded by each client, to be also independent of the bit rate used.

B. The Virtual World Model

We model the virtual world as a square area of certain size. As illustrated in Fig. 1a, we assume that there is no muted avatar and the radii of all the *communication zones* are fixed at 30 m. In another work [5], we have proposed two types of avatar distribution models in the virtual world. The first model is a cluster distribution (Fig. 1c), which captures the popularity locations present in real computer game traces [12]

where gamers converge due to common interest. The second model is the simple uniform distribution (Fig. 1d) where all avatars are randomly placed across the virtual world.

C. The Physical Network Model

The physical network simulated in this work is based on the 5000 node *Transit-Stub* topologies generated by the GT-ITM topology generator [11]. Available server sites were randomly chosen from the 5000 nodes. We also randomly placed 100 ISP POPs across the stubs Autonomous Systems. The maximum node-to-node propagation delay in our topologies was set to 160 milliseconds, based on mean ping-time measurements between Los Angeles, USA and Sydney, Australia...

V. SERVER ASSIGNMENT MODEL

As stated in Section I, in terms of the criteria of voice latency and access upload cost, the two opposing extremes of full-mesh P2P and central server can only achieve the lowest bound on one criterion by incurring (sacrificing) the maximum bound on the other criterion. Consequently, an optimal solution capable of simultaneously achieving both lowest bounds does not exist. The objective of our *Auditory Scene Creation* (ASC) server assignment model is to ascertain a good balance between reducing voice latency and reducing access upload cost. The formulation details are given below in Table I.

TABLE I

PARAMETERS AND VARIABLES FOR THE SERVER ASSIGNMENT PROBLEM

Notation	Meaning
<i>Graph Modelling Terms</i>	
V	The set of all nodes, $V = V_s \cup V_p$
E	The set of edges
G	Graph $G\{V, E\}$
<i>Known Conditions</i>	
V_s	The set of potential ASC server sites.
N_s	The number of ASC server sites $N_s = V_s $
V_p	The set of POPs
N_p	The number of POPs $N_p = V_p $
N_a	The number of avatars. $1 \leq i, j \leq N$
a_{ij}	Indicating if two avatars are communicating $a_{ij} \in [0, 1]$
d_j^s	The two-hop POP-to-server voice transmission delays between the j^{th} avatar and the s^{th} ASC server site.
δ_{ij}	The <i>latency constraint</i> between the i^{th} and the j^{th} avatars.
<i>Decision Variables</i>	
x_i^m	Indicating if the i^{th} avatar uploads its voice stream to the ASC server at the m^{th} site $x_i^m \in [0, 1]$.
y_j^m	Indicating if the ASC server is established for j^{th} avatar at the m^{th} site $y_j^m \in [0, 1]$.
Latency Deviation (%)	percentage difference between the actual latency and the latency constraint for each avatar pair $\left(\frac{d_j^s + d_i^s - \delta_{ij}}{\delta_{ij}} \right) \times 100$ (delay deviations below or at 0% are ignored)
Metric 1 Latency	The percentage of avatar pairs with above zero delay deviation.
Metric 2 Latency	The mean above zero delay deviations incurred by the avatar pairs from Metric 1.

A. Lowest Feasible Delay Server Assignment Formulation

Objective Function:

$$\text{Minimise: } \sum_{m=1}^S \sum_{i=1}^N x_i^m \quad (1)$$

The objective function of *Lowest Feasible Delay Formulation* minimises the sum of uploads from all the avatars to the ASC servers requiring their voices.

Subject to:

$$a_{ij} y_j^m (d_j^m + d_i^m) \leq \delta_{ij} \quad \forall i, j, m \quad (2)$$

$$\sum_{j=1}^N (a_{ij} y_j^m) \leq \left(\sum_{j=1}^N a_{ij} \right) x_i^m \quad \forall i, m \quad (3)$$

$$\sum_{m=1}^S y_j^m = 1 \quad \forall j \quad (4)$$

$$x_i^m, y_j^m \in \{0, 1\} \quad \forall i, j, s$$

The Constraints (2) in the *Lowest Feasible Delay* server assignment Formulation ensures that each communicating avatar pair incurs zero or the lowest *latency deviation* feasible to our two-hop distributed server architecture and the underlying network. The complexity of this formulation is:

- Number of Binary Decision Variables: $2NS$ i.e., $O(NS)$.
- Number of Constraints: $NS+N$ i.e., $O(NS)$.

The *Lowest Feasible Delay Formulation* effectively obtains the lowest bound on *Latency Deviation* at the expense of incurring the maximum bound on access upload cost as returned by the value of Objective Function (1). We verified this maximum bound by executing a simple Facility-Location formulation (*Bandwidth Unconstrained Formulation*) to find the minimum *Latency Deviation* with no constraint on access upload cost. We found that we can not achieve any further reductions in *delay deviation*.

B. Bandwidth Reduced Server Assignment Algorithm

The maximum bound condition of the *Least Feasible Formulation* leads to high access upload cost, especially at high communication density, e.g., 9.77 uploads per avatar at communication density of 30 (at 25 server sites Fig. 4). We developed a *Bandwidth Constrained Formulation* by adding to the *Bandwidth Unconstrained Formulation*, a access bandwidth capacity constraint, e.g., 50% reduction from the maximum bound. Although the *Bandwidth Constrained Formulation* can find the minimum *Latency Deviation* at any bandwidth constraint. We found it to be NP-hard and hence not scalable enough to be of practical use, especially when having to cope with frequent virtual world mobility. Consequently, we have devised a scalable heuristic called *Bandwidth Reduced Algorithm* which iteratively applies the *Lowest Feasible Delay Formulation* until it finds a *sweet spot* of satisfactory trade-off between access upload cost and voice latency. We found only small performance gap between the optimum *Bandwidth Constrained Formulation* and the *Bandwidth Reduced Algorithm*. For the scenario in Fig. 4, at communication density of 30, for the same access upload capacity constraint, the optimum achieved 37.06% of Metric 1 Latency in comparison to the 38.39% achieved by the heuristic, whereas the corresponding numbers for Metric 2 Latency are 30.33% optimum and 31.65% heuristic. In order to reduce the access upload cost, At the k^{th} iteration, the

Bandwidth Reduced Algorithm selects the ASC servers from a reduced set of L_k ($L_k < S$ always) *most delay efficient* sites, rather than from the full set of S available sites. We assume equal leasing cost on all available sites and hence no preferences on which sites are selected to set up ASC servers. In pre-processing, we examine the *Delay Efficiency* of each available site in terms of the number of avatar pairs whose *Lowest Feasible Delay* condition can be met by a central server set up at that particular site. At the k^{th} iteration, L_k is either incremented if the delay deviation of $(k-1)^{th}$ iteration is too high, or decremented if the access upload cost of the $(k-1)^{th}$ iteration is too high. The complexity of the *Lowest Feasible Delay Formulation* applied at the k^{th} iteration is therefore:

- Number of Binary Decision Variables: $2NL_k$, i.e., $O(NL_k)$.
- Number of Constraints: $NL_k + N$ i.e. $O(NL_k)$.

For the 25 sites GT-ITM topology used (see C of Section III), the number of iterations required by the *Bandwidth Reduced Algorithm* to reach the *sweet spot* is 5 or less in all cases studied. In terms of the voice processing load on each server, we assume multiple processors or platforms are available at each site. We observed that the more *Delay Efficient* sites bear heavier processing load. However, the processing reduction scheme proposed by us in [5] can significantly reduce server processing load. In fact, the effectiveness of this scheme increases with the number of avatars assigned to a site.

VI. SIMULATIONS

The GT-ITM physical network topology and the virtual world model applied in our simulations have already been described in details in B and C of Section III. The key parameters to our simulations are as follow:

- Wide spread of 100 POPs (N_p) and 25 potential server sites (N_s). No correlation between the physical world and the virtual world.
- 2000 avatars of cluster distribution and communication density of 30.
- The simulation results were averaged over 100 runs, e.g., 20 GT-ITM topologies and 5 random sets of 25 sites per topology.

A. Bandwidth Reduction Performance of the Bandwidth Reduced Algorithm

Fig. 4 illustrates the iterative process (from 25 to 4 sites) of the *Bandwidth Reduced Algorithm* to find a *sweet spot* of good trade-off between the access upload cost and delay performance. This *sweet spot* occurs when 5 *Delay Efficient* sites are used. At this *sweet spot*, access upload cost is reduced from 9.77 to 3.41 streams per avatar at the expense of 5.77% increase in Metric 1 Latency and 4.53% increase in Metric 2 Latency, in comparison to when the full set of 25 sites is applied. Additional analysis reveals that only relatively small delay reductions can be obtained when increasing the full set of available sites from 25 to 50 sites for 100 POPs. This agrees with Nguyen's findings [10].

B. Effect of Variations in the Avatar Population

In this section, we analyse the impact of varying avatar distribution on the performance of our server assignment solutions. Avatar distribution in the virtual world can be varied according to three main parameters, i.e., communication density, avatar number and the types of avatar distribution (uniform and cluster distributions).

Figure 4. Iterative process to find a sweet spot between bandwidth and latency

In terms of *Latency Deviation*, the effect of rising communication densities is not very significant. For example, in terms of the Metric 1 Latency achieved by the *Least Feasible Delay Formulation*, there is less than 10% rise (from 31.01% to 39.23%) when the communication density increases 5 times from 10 to 50. We investigate uncorrelated DVEs such that avatars in the same communication zone are likely to access very different POPs in the physical world. In addition, because of the fixed locations of POPs (local exchanges), our server assignment solution can not influence the client-to-POP access link delay. These two factors explain the lack of dependency between communication density and the delay performance gap between different architectures. For the same reasons, the variations in the number of avatars and types of avatar distributions also have negligible effect on the delay performance of different architectures. This finding agrees with our results in voice processing cost reduction for Immersive VoIP service in [5]. On the other hand, when communication density is increased, both the full-mesh P2P and the *Least Feasible Delay Formulation* incur significant increase in access upload costs whereas the *Bandwidth Reduced Algorithm* incurs very small rise in access upload cost.

C. Impact of Avatar Mobility on Server Assignment

In this section, we investigate the impact of avatar mobility in the virtual world on the execution frequency of the *Bandwidth Reduced* server assignment algorithm. To simulate the mobility of avatars in the virtual world, we applied the Networked Game Mobility Model (NGMM) [12] which modifies the definition of stationary states in the Random Waypoint Motion (RWM) model. We simulated the speed range of 0.1 to 5 meters per second. We set 0.5 second of pause time between the stationary and mobile states to simulate a fast paced First-Person-Shooter (FPS) game. We

studied the *Latency Deviation* comparison between the case of reusing the server assignment solutions from $t=0$ seconds at subsequent instants and the case of server re-assignments every 10 seconds. We define the Metric 1 Penalty as the additional Metric 1 Latency incurred in the case of no server reassignment over the case of server reassignment. Similarly, the Metric 2 Penalty is the additional Metric 2 Latency incurred due to no server reassignment. For the mobility model studied, we chose the execution frequency of the *Bandwidth Reduced Algorithm* so as to never permit the deterioration in delay performance to exceed 1/4 of the delay improvement achieved by the *Least Feasible Delay Formulation* over central server. This translates to 4.5% of Metric 1 and/or Metric 2 Penalties. For the mobility model studied, an execution frequency of once per 80 seconds is required for the low communication density of 5 which is probably best served by P2P rather than a client/server architecture which can not cope due to the delay server handover and reassignment. As the communication density rise, the impact of avatar mobility diminishes noticeably. At both avatar densities of 15 and 25, the execution frequency can be set at or beyond 1000 seconds or 16 minutes which is also ample time for the *Bandwidth Reduced Algorithm* to re-execute (see Table V) and the server handover mechanism to take place. Although not shown, the additional bandwidth cost incurred by not regularly performing server reassignment increase very slightly with time, e.g., from 0 stream per avatar at 10th second to 0.8 stream per avatar at 1000th second, at communication density of 25. This deterioration effect is even more negligible at lower avatar densities.

D. Algorithm Scalability

We implemented the *Voice Processing Minimisation Algorithm* using the CPLEX software package [14] on a PC running Linux with AMD Athlon64 Dual-Core 2.0 GHz processor and 4.0 GB of RAM. As shown in Table II, both the simplified *Lowest Feasible Delay* formulation and the *Bandwidth Reduced Algorithm* can scale reasonably well up to 20,000 avatars. This is enough for a current *shard* in a MMOG such as World of Warcraft. As a future work, in order to support beyond 20,000 avatars, we plan to separate the large virtual world into partitions and solve independently (parallelisation) the server assignment problem in each partition. This partitioning scheme combined with a scheme of replacing overloaded (processing) sites will help address the processing overload problems at certain *Delay Efficient* site.

VII. CONCLUSION

In this work, our subjective listening results verified that our conjecture of variable latency constraints governed by virtual world distance. We found this conjecture to be more perceptually accurate and network feasible than fixed latency constraint. We have also shown that based on the bounds found by the *Lowest Feasible Delay Formulation*, the *Bandwidth Reduced Algorithm* can offer a good trade-off between reducing latency and reducing access upload cost, even in the challenging scenario of no clear correlations between the DVE virtual world and the physical network. Moreover, we have also found that neither the variation of avatar distribution nor avatar mobility can have any significant impact on our server assignment solutions. Both

the *Lowest Feasible Delay Formulation* and the *Bandwidth Reduced Algorithm* have reasonable scalability. Hence we conclude, for the mobility model studied, our server assignment solution can cope with avatar mobility well.

TABLE II
EXECUTION TIME OF SERVER ASSIGNMENT SOLUTIONS

Number of avatars (At Communication density of 50)	Execution time of Lowest Feasible Delay Formulation (1 iteration)	Execution time of Bandwidth Reduced Algorithm (5 iterations)
200	0.02 seconds	0.07 seconds
2000	6.78 seconds	22.45 seconds
20000	217.89 seconds	402.68 seconds
Avatar densities (For 2000 avatars)	Execution time for Lowest Feasible Delay Formulation (1 iteration)	Execution time (milliseconds) Bandwidth Reduced Algorithm (5 iterations)
10	1.2 seconds	3.00 seconds
30	4 seconds	10.00 seconds
50	6.78 seconds	16.95 seconds

ACKNOWLEDGEMENT

This work is supported by the Co-operative Research Centre for Smart Internet CRC (SITCRC) and the University of Wollongong (UOW), Australia.

REFERENCES

- [1] Microsoft Corporation, Xbox LIVE, <http://www.xbox.com/en-US/live>, (16 May 2007).
- [2] MMOGCHART.com, <http://www.mmogchart.com/> (10 April 2007).
- [3] Kuan-Ta Chen and Chi-Laung Lei, "Network Game Design: Hints and Implications of Player Interaction", in *Proc. of 5th ACM workshop On Network & System support for Games (NetGames 06)*, Singapore, 29/Oct.-2/Nov., 2006.
- [4] Hew, K., Gibbs, M. R., and Wadley, G., "Usability and sociability of Xbox Live voice channel", In *Proc. of Australian Workshop on Interactive Entertainment (IE2004)*, pp. 51-58.
- [5] Ying Peng Que, Paul Boustead, and Farzad Safaei, "Voice Processing Reduction in an Immersive Voice Over IP Service to Wireless Gaming", in *Proc. of the 6th Annual Workshop on Network and Systems Support for Games (NetGames 06)*, Melbourne, Australia, 19-20 September 2007.
- [6] E. C. Cherry, "Some further experiments upon recognition of speech with one and with two ears", in *Journal of Acoustics Society America*, Vol. 28, pp. 889-895, 1956.
- [7] Ying Peng Que, Paul Boustead, and Farzad Safaei, "Scalable Delivery of Immersive Communication Environment (ICE) using Peer-to-Peer (P2P) Architecture", in *Proc. of the 49th Annual IEEE Globecom Conference (GLOBECOM 2006)*, San Francisco, 27 Nov.-1 Dec. 2006.
- [8] G. Scheets, R. Singh and M. Parperis, "Analysing End-to-End Delivery Delay in Pure VoIP Networks", in *45th Midwest Symposium on Circuits and Systems*, 2002, vol. 2, pp. 517-520, Aug. 2002.
- [9] Elizabeth M. Wenzel, "Analysis of the role of update rate and system latency in interactive virtual acoustic environments", in *Proc. the 103rd Audio Engineering Society Convention*, preprint 4633, Sept. 1997.
- [10] C. D. Nguyen, F. Safaei, P. Boustead, "A distributed server architecture for providing immersive audio communication to massively multi-player online games", in *Proc. of IEEE International Conference on Networks (ICON)*, Singapore, November 2004, pp. 170-176.
- [11] K. Calvert, M Doar, and E. Zegura, "Modelling Internet Topology", in *IEEE Communications Magazine*, vol. 35, pp. 160-163, June 1997.
- [12] Tan S.W., Lau W. and Loh A., "Networked Game Mobility Model for First-Person-Shoot Games", in *Proc. of the 4th ACM workshop On Network & System Support for Games (NetGames 05)*, Hawthorn, NY, USA, 10-11 Oct. 2005.
- [13] Methods for subjective determination of transmission quality, International Telecommunications Union (ITU), Geneva, Switzerland, ITU-T Recommendation P.800, August, 1996.
- [14] Cplex optimisation software, documentation available at <http://www.ilog.com>, (9 Jan. 2007).
- [15] Paul Boustead and Farzad Safaei, "Comparison of Delivery Architectures for Immersive Audio in Crowded Networked Games", in *Proc. of the 14th ACM International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV)*, Ireland, 16-18 June, 2004.
- [16] Cisco Systems Inc, Understanding Delay in Packet Voice Networks, White Paper, Document ID: 5125, <http://www.cisco.com/warp/public/788/voip/delay-details.html>.