Demo: DeLorean: Using Speculation to Enable Low-Latency Continuous Interaction for Mobile Cloud Gaming

Kyungmin Lee* David Chu† Eduardo Cuervo† Johannes Kopf† Alec Wolman† Jason Flinn*
*Univeristy of Michigan †Microsoft Research

Playing games on mobile devices is very popular. Recently, cloud gaming – where datacenter servers execute the games on behalf of thin clients that merely transmit UI input events and display output rendered by the servers – has emerged as an interesting alternative to traditional client-side game execution. Cloud Gaming-as-a-Service (GaaS) offers several advantages salient to mobile clients. First, users with low end devices can get the same high quality experience as users with high end devices. Second, mobile game developers avoid two challenges that arise with the huge diversity of mobile devices: platform compatibility headaches and per-platform performance tuning. Third, upgrading servers (e.g., for bug fixes, game updates, etc.) becomes far easier than redeploying new software to clients. Finally, players can select from a vast library of games and instantly play any of them.

However, GaaS on mobile devices faces a key technical dilemma: how can players attain *real-time interactivity* in the face of wide-area latency? Real-time interactivity means client input events should be quickly reflected on the client display. User studies have shown that players are sensitive to as little as 60 ms latency, and are aggravated at latencies in excess of 100 ms [1]. A further delay degradation from 150 ms to 250 ms lowers user engagement by 75% [2].

Instead, we propose to mitigate wide-area latency via speculative execution. We present *DeLorean* a system that delivers real-time gaming interactivity as fast as traditional local client-side execution, despite with network latencies.

DeLorean's basic approach combines input prediction with speculative execution to render mulitple possible frame outputs which could occur RTT milliseconds in the future. DeLorean employs the following techniques to accomplish this.

Future Input Prediction: Given the user's historical tendencies and recent behavior, we show that some categories of user actions are highly predictable. We develop a Markov-based prediction model that examines recent user input to forecast expected future input. We use two techniques to improve prediction quality: supersampling of input events, and constructing a Kalman filter to improve users' perception of smoothness.

State Space Subsampling and Time Shifting: Certain user inputs (e.g., firing a gun) cannot be easily predicted. For these, we use parallel speculative executions to explore multiple outcomes. However, the set of all possible frames over long RTTs can be very large due to state space explosion. To address this, we use two techniques: state space subsampling, and event stream time shifting. These greatly reduce possible outcomes with minimal impact on the quality of interaction, thereby permitting speculation within a reasonable budget.

Misprediction Compensation: When mispredictions occur, DeLorean enables the client to execute *error compensation* on the (mis)predicted frame. The resulting frame is very close to what the client ought to see. Our misprediction compensation uses *view interpolation*, a vision technique that transforms pre-rendered images from one viewpoint to a different viewpoint using only a small amount of additional 3D metadata.

To punctuate our emphasis on fast interaction, we evaluate DeLorean's prediction techniques using two *fast action games* where even small latencies are disadvantageous. Doom 3 is a twitch-based first person shooter where responsiveness is paramount. Fable 3 is a role playing game with frequent fast action combat. Both are high-quality, commercially-released games, and are very similar to mobile games in the first person shooter and role playing genres, respectively.

Through interactive gamer testing, we found that players perceived only minor differences in responsiveness on DeLorean even with some network latnecy when compared head-to-head to a system with no latency. Overall, player surveys indicated positive reception of gameplay on DeLorean.

Categories and Subject Descriptors

D.2.2 [**Software Engineering**]: Design Tools and Techniques

Keywords

Cloud gaming; Speculation

MobiSys'14, June 16–19, 2014, Bretton Woods, New Hampshire, USA.
ACM 978-1-4503-2793-0/14/06.
http://dx.doi.org/10.1145/2594368.2601474.

1. REFERENCES

[1] Beigbeder, T., Coughlan, R., Lusher, C., Plunkett, J., Agu, E., and Claypool, M. The effects of loss and latency on user performance in unreal tournament 2003. In *NetGames'04* (New York, NY, USA, 2004), ACM, pp. 144–151.

[2] Chen, K.-T., Huang, P., and Lei, C.-L. How sensitive are online gamers to network quality? *Commun. ACM 49*, 11 (Nov. 2006), 34–38.