Towards Zero-Latency User Experiences

Ricardo Jota, Clifton Forlines, Darren Leigh, Steven Sanders, Daniel Wigdor

Tactual Labs Co. | www.tactuallabs.com | info@tactuallabs.com

ABSTRACT

To understand why we researched latency and sought zero-latency as a goal, we must first define the term. Our research addresses "overall latency" as the time between the user's finger arriving on the touchscreen, to the time the system's response to that input is displayed. Once latency is understood, along with its historical measurement and persisting issues, the reader will understand why solving the latency problem holds exciting potential for users and developers across many industries and applications.

Keywords: touch input, multi-touch, user input, latency

OTHER SOURCES

This whitepaper summarizes and extends our team's research published in leading venues of the Association for Computing Machinery. Please see original publications for more detailed descriptions of system design [2], as well as user perception of [1,2] and performance under latency [1].

INTRODUCTION

Touch-driven interfaces are ubiquitous today, most commonly found in mobile phones, tablets, laptops, e-readers, and larger interactive surfaces. Despite industry shift to capacitive touchscreens, a persistent problem is latency—the time between a finger touch and the on-screen response. Latency has been identified as an important issue across interactive systems and indirect input devices [6, 7, 9].

Today's touchscreen devices commonly exhibit latency between 50 and 200ms. Such delay is especially noticeable when users interact with gaming applications, graphics tools, or any task where they move objects around a screen [1]. Graphical tricks have been shown to reduce users' perception of latency [11], but are unable to mask it entirely. Further, those perceptual tricks still limit user performance of basic tasks required to operate interactive systems.

Scientific publications have long held that a response within 100ms (1/10th of a second) is sufficient to be imperceptible. This is largely based on research by Miller, published in 1968: "response should be immediate and perceived as part of the mechanical action induced by the operator. *Time delay:* No more than 0.1 second(s)" [6].

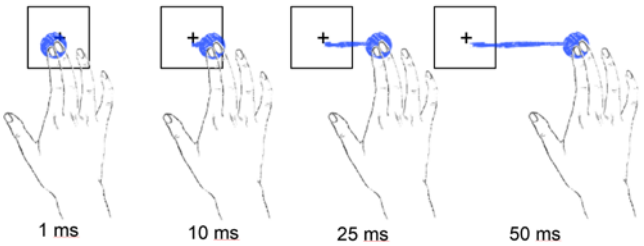

Figure 1: Latency causes physical separation between an on-screen object and the finger dragging it. (Tether added for emphasis.)

This 100ms figure has evolved as the benchmark for 'instantaneous' response. Cited in journals, patents, government procurement guidelines, university classes, and more, such benchmark has long-stood. What is critical, however, is the underlying experiments were limited by technology of the time. In particular, the devices used were *indirect* – i.e., input location and on-screen response were physically separated. In addition, operations were limited to discrete input actions such as clicking a button, rather than continuous actions such as dragging items around a screen.

In contrast, modern direct touch systems show the system's response to users' input is directly alongside their finger. Only that can create a zero-latency representation of their input and highlight differences between input and latent output locations. As Figure 1 illustrates, combined with the continuous operations of modern systems, latency is very obvious to the user.

Effects of Latency

Our research was designed to understand effects of latency on user interaction with direct touch systems. In particular, we sought answers to two questions:

1. *What level of latency is perceptible to users of direct-touch systems?*
2. *What is the effect of latency on users' abilities to perform everyday tasks with direct-touch systems?*

Human Perception

How much latency can a user really notice when engaging with an interactive system? Prior attempts to answer were null, as none were equipped with a testing apparatus capable of under 1ms of latency. So,real-time interaction was impossible to evaluate. Our work examined users' ability to perceive latency of system responses to two input types: tapping input (clicking a button) and dragging input (scrolling a list or moving an icon). As described later, we concluded that users are able to perceive far lower degrees of latency than previously believed. With practice, users could perceive less than 5ms of latency.

Human Performance

While users might be able to *perceive* lesser degrees of latency, does latency actually impair users' abilities to *perform* everyday tasks? Does it take longer to scroll a list, move an icon, line-up an "Angry Bird" launch, or adjust a car's controls when subjected to greater latency? Our findings surprised: even extremely low levels of latency caused reduced performance in everyday tasks. As little as 25ms was enough to impair performance.

Before presenting those studies in greater detail, we examine the high levels of latency in today's devices and their sources.

LATENCY IN TODAY'S DEVICES

TouchMarks Latency Measurements

The *Touchscope* is a simple apparatus that measures overall latency. Agawi developed and used the device to measure latency of modern touch tablets. The results, shown in Figure 2, confirm that a very high degree of latency exists in even top-of-the-line hardware. The iPad Mini surfaced as best-in-class, with a mean latency of 75ms. While better than other consumer devices, 75ms latency remains an eternity, leading to a noticeable lag that negatively impacts user performance. Those results, alone, explain the veritable breakthrough of our 1ms touch system.

Sources of Latency

Latency in a direct-touch device has many sources, usually with three main components: 1) the physical sensor that captures touch events; 2) the software that processes touch events and generates output for the display; 3) the display itself. Reducing system latency in a direct-touch system required addressing latency issues in all three components, a significant challenge. We therefore adopted a holistic, systems approach to what has been traditionally viewed as the domain of componentry.

Figure 3 shows the flow of information from time of user touch to the screen, along with approximate time required by each system component. Note the large range of possible times for processing. This mostly reflects architectural issues, differences between operating systems, and the efficiency of application design.

We believe these longstanding sources of latency can be eliminated. We developed our own prototype system capable of generating responses to user input in less than 400µs (0.4ms). This system enabled us to run experiments examining user perception of latency, as well as its effects on performance of normal computing tasks.

PERCEPTION OF LATENCY

While earlier work suggested users are able to perceive latency down to 100ms, our experiments made it clear that, for direct-touch systems, far lower levels of latency were perceptible. To guide our work in reducing latency, we needed

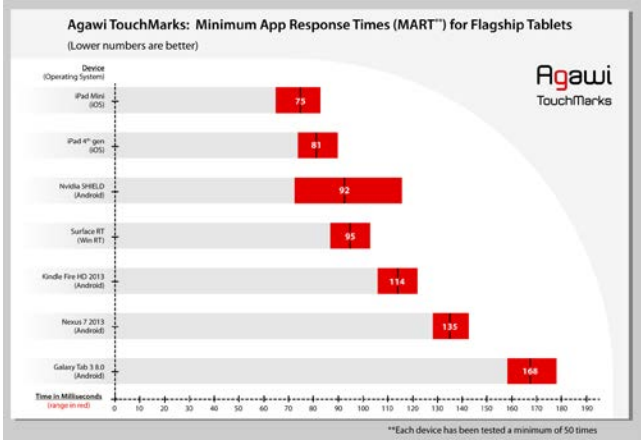

Figure 2. Agawi's measurements of overall latency of modern touch devices (from: http://ap-pglimpse.com/blog/)

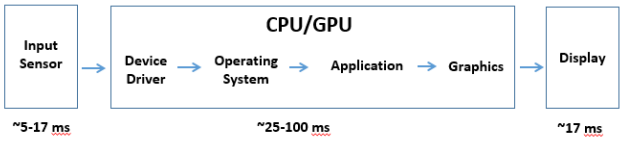

Figure 3: Sources of latency in modern computer systems.

to know at precisely what latency value users would no longer notice a difference. Thus, we conducted a formal experiment to modulate overall system latency, and measured participants' ability to perceive the differences in performance.

Just Noticeable Difference

We conducted experiments to determine the just-noticeable difference (JND) of various performance levels. JND is defined as the threshold level at which a participant is able to discriminate between two unequal stimuli – one consistently presented at the same level (the reference), and one whose value is changed dynamically throughout the experiment (the probe) [4]. A commonly accepted value for JND at some arbitrary reference value is a probe at which a participant can correctly identify the reference 75% of the time [4]. A probe value that cannot be distinguished from the reference with this level of accuracy is considered 'not noticeably different' from the reference.

Participants

Ten right-handed participants aged 24-40 were recruited from the local community and paid $20 for an approximately 1-hour study.

Procedure

Participants were repeatedly presented with pairs of latency conditions: the reference value (1ms latency) and the probe (between 1 and 65ms). Participants dragged their finger from left to right, then right to left on the touchscreen display. Beneath the user's contact point, the system rendered a solid white 2cm × 2cm square. The speed of movement was left up to the participants.

The order of the conditions was randomized for each pair. The study was designed as a two-alternative, forced-choice experiment, i.e., participants were instructed to choose, within each trial, which case was the reference (1ms) value and were not permitted to make a "don't know" or "unsure" selection [9]. After each pair, participants informed the experimenter which of the two was "faster".

Design

For each trial to converge at our desired 75% JND level, the amount of added latency was controlled according to an adaptive staircase algorithm, a more recent and now broadly adopted approach than previous staircase and modified staircase algorithms [4].

Each correct identification of the reference value caused a decrease in the amount of latency in the probe, while each incorrect response caused the probe's latency to increase. To

reach the 75 % confidence level, increases and decreases followed the simple weighted up-down method described by Kaernbach, wherein increases had a three-fold multiplier applied to the base step size, and decreases were the base step size (initially 8ms) [1]. When a participant responded incorrectly after a correct response, or correctly after an incorrect response, this was termed a reversal, as it caused the direction of the staircase (increasing or decreasing) to reverse. The step size, initially 8ms, was halved at each reversal, to a minimum step size of 1ms. This continued until a total of 10 reversals occurred, resulting in a convergence at 75% correctness.

Each participant completed eight staircase 'runs'. Four started at the minimum probe latency (1ms) and four at the maximum (65ms). The higher starting value of the staircase was chosen because it roughly coincides with the best commercial offerings, and because pilot testing showed this value would be differentiated from the 1ms reference with near 100% accuracy, avoiding ceiling effects.

Staircases were run two at a time in interleaved pairs to prevent response biases otherwise caused by participants' ability to track progress between successive stimuli [4]. Staircase conditions for each pair were selected at random without replacement from 8 possibilities (2 starting levels × 4 repetitions). The entire experiment, including breaks between staircases, was completed by each participant within a single 1-hour session.

Results

Participant JND levels ranged from 2.38ms to 11.36ms, with a mean JND across all participants of 6.04ms (standard deviation 4.33ms). JND levels did not vary significantly across the 8 runs for each participant. Figure 4 shows results for each participant.

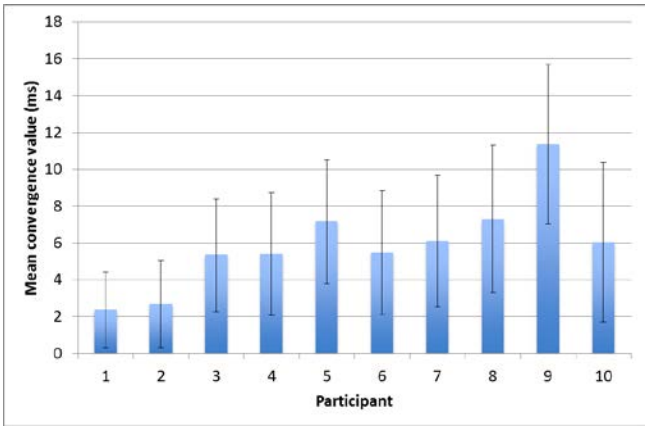

Figure 5: Mean perceptible latency, per participant. Overall mean was 6ms, an order of magnitude (10X) faster than best devices available today.

Discussion

Participants were able to discern differences in latency far below the threshold of current consumer devices. Our results confirmed that an order of magnitude (10X) improvement in latency would be noticed and appreciated by users of touch devices.

PERFORMANCE UNDER LATENCY

To test performance impacts, we asked participants to repeatedly perform pointing tasks on a touch screen display, and included *latency* among the experiment's design factors. This allowed us to observe the effect of latency on performance, and if there was an effect, to determine how users changed behavior to account for latency.

Participants

Forty-five participants from the local community, ages 19 to 52, took part in the study. Each was paid $10 for a one-hour session. All were right-handed and had experience using touch devices.

Task

Interaction with modern computing systems, while comprised of complex user interfaces and gestures, can be abstracted to a small set of basic tasks. ISO 9241-9 defines the one-direction tapping task [4] as one of the fundamental tasks of computing: pointing and clicking. Modern touch devices are similar, with one exception: content is *dragged* (e.g., when scrolling a list or moving an object from one place to another). Thus, we modified this task to require the participant to touch a point on the screen, then drag their finger across it. The task is described in Figure 5.

Procedure

The procedure followed ISO9241-9, whereby participants were asked to maintain an error rate of 5% by either speeding up or slowing down. This was to maintain consistency across participants and to avoid confounds with the speed-accuracy tradeoff.

Design

Dragging tasks varied according to three independent variables: *latency* of the cursor movement (1, 10, 25, and 50ms artificially inserted between input frames); *width* of the target (3, 4, and 5cm; the cursor is a box that measured 2cm as in [2] and *distance* between starting position and target (3.5, 8.5, and 15cm).

Each participant performed 8 repetitions of all 36 combinations of levels of *latency*, *target size*, and *distance*, for a total of 288 trials per participant. Order of the 288 trials was randomized across participants. In summary, the design of the experiment included:

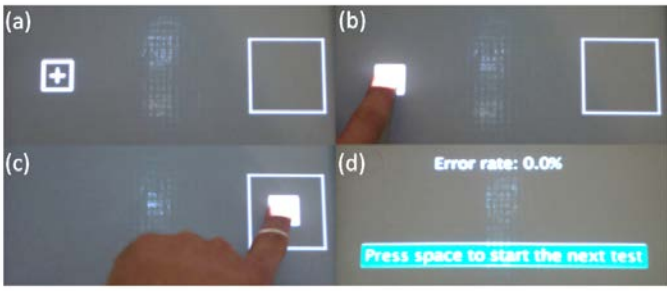

Figure 4: The dragging task, modeled after ISO9241-9: (a) before user selects the cursor; (b) after the cursor is selected; (c) cursor has been dragged to the target; (d) user lifts their finger.

4 levels of latency
3 target sizes
3 target distances
8 repetitions
x 45 participants
= 12,960 total trials

RESULTS

Results are grouped by hypothesis, after which we examine effects of latency on different phases of the task. We conclude with the more subtle effects of latency on pointing using a direct-touch input device.

Analysis of Performance Measures

Figure 6 shows that latency had a significant effect on time required to complete the task ($F_{3,132} = 45.128$, $p < 0.001$). In particular, we found a significant effect for the interaction between latency and target size on error rate ($F_{6, 264}=2.581$, $p=0.019$) and movement time ($F_{6,264}=5.782$, $p<0.001$). Figure 6 also highlights the significant effect for the interaction of distance and latency on movement time ($F_{6,264}=3.483$, $p=0.02$). These results indicate that the effects of latency on performance are more pronounced for small targets, and for targets that are farther away.

These results clearly confirm that latency impairs user performance of basic tasks. To understand the degree of impairment, we conducted *post hoc* pairwise comparisons of particular latency values. These revealed no significant difference of the effects on throughput between the lowest two tested latencies (1ms and 10ms). Importantly, all other pairs showed significant differences.

This finding suggests a possible floor for the benefit of reduced latency at 10ms. It appears that below 10ms, further reductions in latency may not reduce impairment. However, our results remain inconclusive on this point. We performed a linear regression on the results of movement time for the latencies of 10ms, 25ms, and 50ms, omitting the lowest latency value. Figure 6 shows that this line fits the results precisely ($R^2=0.956$). Using this line to predict the mean movement time for latency=1ms, we find an expected mean of 747.6ms. Our empirical results report a mean of 751.3ms, well within the 95% confidence interval predicted. This evidence suggests there may not be a performance floor. Regardless, how low latency levels become, any latency at all in the system may continue to impact user performance.

While we are unable to conclude whether 10ms is in fact a floor (i.e., that it is "fast enough" to eliminate latency-caused impairment of user performance), our results confirm that current latency performance of consumer devices, summarized in Figure 2, is sufficient to have a significant effect on user ability to perform everyday computing tasks. As shown in Figure 6, eliminating input latency has potential to improve speed with which users perform input to touch screens by more than 10%. Considering the number of scrolling and other gestures users perform while engaged with a touch screen over the span of a usage session, this represents significant time savings.

ACKNOWLEDGEMENTS

We thank the members of the Dynamic Graphics Project at the University of Toronto, and the Applied Sciences group at Microsoft for valuable assistance.

REFERENCES

1. Jota, R., Ng, A., Dietz, P., & Wigdor, D. (2013, April). How fast is fast enough?: a study of the effects of latency in direct-touch pointing tasks. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2291-2300). ACM.
2. Ng, A., Lepinski, J., Wigdor, D., Sanders, S., & Dietz, P. (2012, October). Designing for low-latency direct-touch input. In Proceedings of the 25th annual ACM symposium on User interface software and technology (pp. 453-464). ACM.
3. Kaernbach, C. 1991. Simple adaptive testing with the weighted up-down method, *Perception & Psychophysics* 49, 227-229.
4. ISO/DIS9241–9 Ergonomic requirements for office work with visual display terminals (VDTs) – Part 9. International Standard, International Organization for Standardization, 2000.
5. Levitt, H. 1971. Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, 49(2): 467–477.
6. MacKenzie, I. S., and Ware, C. (1993). Lag as a determinant of human performance in interactive systems. ACM CHI 1993.
7. Meehan, M., Razzaque, S., Whitton, M. C., and Brooks, F. P. (2003). Effect of Latency on Presence in Stressful Virtual Environments. *IEEE VR* 2003, 141–138.
8. Miller, R. B. 1968. Response time in man-computer conversational transactions. *Fall Joint Comp. Conf.* 1968, 267-277.
9. Savage, W. C. 1970. *The Measurement of Sensation: A Critique of Perceptual Psychophysics*. Berkeley: UC Press.
10. Steed A. (2008). A simple method for estimating the latency of interactive, real-time graphics simulations. In *Proceedings of VRST '08*. ACM, New York, NY, USA, 123-129.
11. Wigdor D., Williams S., Cronin M., Levy R., White K., Mazeev M., and Benko H. (2009). Ripples: utilizing per-contact visualizations to improve user interaction with touch displays. In *Proceedings of UIST '09*. ACM, New York, NY, USA, 3-12.

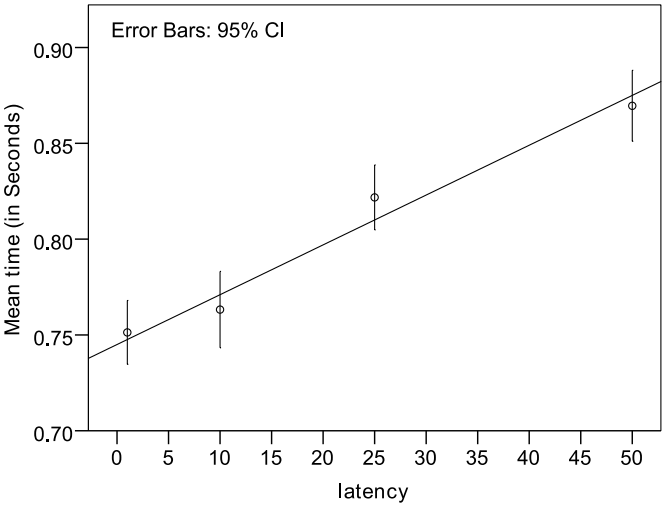

Figure 6: Results of mean time required to perform everyday pointing tasks under different levels of latency. Regression line predicts continuous benefit of decreased latency. Note cut vertical axis.