Speak<geek>
Tech Brief

1

RichRelevance Infrastructure: a robust, retail-optimized foundation

rr richrelevance®

RichRelevance Infrastructure: a robust, retail-optimized foundation

Internet powerhouses Google, Microsoft and Amazon may not see eye-to-eye on much, but they do agree on one thing: in the online environment, business is gained (or lost) with every millisecond of response time. In the online retail environment, speed is an even stronger currency. A shopper's perception of page load time can significantly affect shopping behavior and thus decrease revenue. Amazon found that every 100 ms increase in load time equated to a decrease in sales of 1% (Kohavi & Longbotham 2007).

As a result, the stability and performance of any server infrastructure associated with an ecommerce site is of utmost importance. For most online merchants, this includes any third party solutions implemented on their sites (such as recommendations, ratings and reviews), which are supported by outside server environments. The slightest page load delay caused by a slow vendor solution can result in a customer bouncing off a site – permanently. This Speak Geek paper will discuss the importance of choosing vendors like RichRelevance that have doubled down on their infrastructure, architecting cutting edge systems that remain unaffected no matter the level of a retailer's site traffic. For every RichRelevance recommendation product, response times remain consistently under 100 ms with the vast majority of requests leaving RichRelevance data centers in less than 25 ms.

Award-Winning* Capacity and Speed

RichRelevance delivers recommendations to shoppers via its SaaS platform. These are based off of real-time data collected about user behavior, catalog data, and inventory status combined with historical user and product information and trends. In order to keep each recommendation relevant, recommendation models are rebuilt several times a day based on complex mathematical models. To manage such a tremendous amount of data, RichRelevance uses a cloud-computing model organized over six geographically diverse data centers equipped with load balanced tier 1 servers using solid state disks. This system provides the highest possible throughput with the lowest latency.

This approach means that large spikes in traffic are handled easily without affecting performance. During the peak of 2008

RichRelevance's infrastructure is so cutting-edge that the {rr} IT team, led by Elya Kurktchi, received the 2009 Intel Premier IT Knowledge Award for use of SSDs (solid state drives) in SaaS architecture.

holiday traffic, RichRelevance operated at less than one-fifth of capacity despite serving 1.2 billion product recommendations with approximately 300 million page views and recommendation clicks (between Thanksgiving and Cyber Monday). If any RichRelevance e-commerce client were to experience even a 1000% spike in traffic, all other client sites would continue with business as usual.

Each data center utilizes seven to 10 bandwidth providers to enable the system to pick the fastest provider on a per user basis to deliver each recommendation. The system is also architected in such a way that a user gets his/her request served from the fastest responding data center available.

To further enhance the system's speed and efficiency, RichRelevance installed SSDs (solid state drives) that slash page load times and enable the system to simultaneously run more complex, data intensive recommendation algorithms. With such algorithms, data not available in cache often needs to be accessed. With an SSD, this can be accessed with performance comparable to RAM, up to 1000 times faster than ordinary hard drives.

Six Data Centers Across the US and Growing

- San Francisco, CA
- San Diego, CA
- Dallas, TX
- Houston, TX
- McLean, VA
- New York, NY

How Latency Really Adds Up:

{rr} Multiple Data Centers Approach

25ms	Average round trip time to nearest {rr} data center
x 2	Two round trips to establish a connection on first page hit
= 50ms	Network time for first request
+ 5ms to 25ms	Server response time with with SSD optimized recommendation servers
= Latency of 55ms (average) to 75ms (99th percentile)	

Single Data Center Approach

80ms	Average round trip time from entire US to a single data center
x 2	Two round trips to establish a connection on first page hit.
= 160ms	Network time for first request
+ 20ms to 150ms	Server response time with moderately optimized infrastructure
= Latency of 180ms (average) to 310ms (99th percentile)	

Solid State Drives – Faster and, well ... Faster

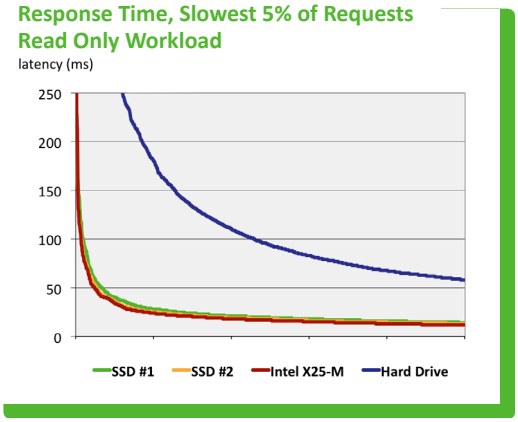
When researching SSD technology and its possible applications in the late summer of 2008, RichRelevance found that although the technology had great promise, not all SSDs were equally useful. All solid state drives performed well under a load of pure random reads. But when writes occured to the drive at the same time as reads, performance diminished significantly on some SSDs — to the point that reads during these writes were worse than a hard drive with the same workload. One SSD stood out from the others, a drive that had just hit the market, the Intel X25-M SSD. Writes of all types performed well but, in particular, concurrent writes did not stop the SSD from performing reads at a high level at the same time. In RichRelevance tests, there was no condition in which it performed worse than a normal hard drive.

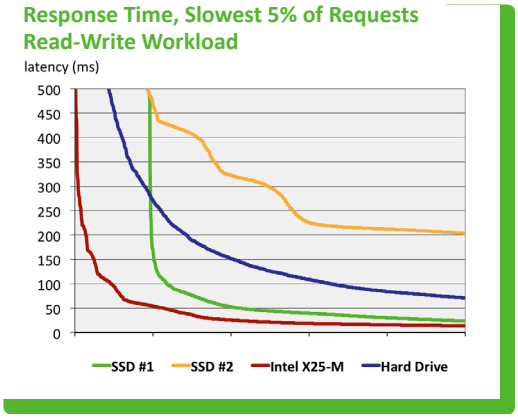
RichRelevance's primary stress test was to run our application at over 20 times the typical load on a single server and measure the response time of recommendations. While this was occurring, we would push an update to our entire dataset. This push mimicked the write load that occurs in RichRelevance's actual production environment when we update our data every hour or so.

The two graphs to the right demonstrate the performance of the SSDs versus a hard drive in the two scenarios. In the Read Only scenario all SSDs perform excellently. In the second scenario we push an update to the drive that causes significant disk write activity during the test. The second graph demonstrates a histogram plot of the 5% of requests with the highest (slowest) latency. 95% of requests were lower (faster) latency.

In the Read Only workload, all SSDs tested perform significantly better than the hard drive. When concurrent writes are introduced, the picture changes dramatically. The Intel SSD slows down with the 99th percentile of requests going from about 25ms to 55ms. The hard drive's 99th percentile goes from 160ms to 225ms. Both of the other SSD's 99th percentile climbs from below 30ms in the read only test to above 500ms in the Read Write test.

These load tests, which were significantly more stressful than ordinary production loads, proved to us that the right SSD technology can be a game-changer.





No Single Point of Failure

To further reinforce its cutting edge usage of new technologies, RichRelevance's IT solution does not rely on access to a centralized data repository to respond to requests for recommendation content. As a result, any single data center failure will not affect recommendations or the user experience. Because each data center has a constantly updated, local cache, individual recommendation requests do not depend on communication with other data centers. This distributed architecture safeguards against downtime, setting RichRelevance apart from other SaaS vendors who rely on centralized infrastructure.

Raising the Bar

When evaluating a third party SaaS vendor, taking a careful look at server environment is crucial. An investment in a vendor with high quality infrastructure supports fundamental site attributes in today's highly competitive environment and ensures a problem-free customer shopping experience while protecting the bottom-line. Any weakness in a provider's infrastructure will be directly reflected on your site and your brand. Slow load times, impaired functionality, and even downtime can be the result of a faulty provider.

Confirming the material importance of a robust server environment, Google found that the long-term effect of poor performance on user visits was significant. Users that experienced reliable and fast load times were more likely to return to a site. As today's retailer knows, increasing the rate of retention is crucial in building lifetime customer value (LTV). Thus, site speed and reliability are inextricably tied to a gold standard in measuring performance and future success.

Check www.richrelevance.com/speekgeek for the latest downloadable edition in this series.

Quick Fixes In-House:

What Merchants Can Do To Improve Page Load Speed

- Build pages so that rendering can begin before all of the page is downloaded.
- Optimize the order that first party and third party components are called:
  - Prioritize the most critical data and fast components (such as RichRelevance products) so that they are visible to the user first regardless of the status of other components.
  - Allow components that can download asynchronously to do so as early as possible.
- Ensure that cacheable resources are configured to utilize browser caching.
- Take advantage of content compression on all valid resources from all sources — that includes all text!