Scaling

* If we want to start writing apps for thousands, or perhaps even millions, of users then obviously we'll need to think about things a little differently.
* This process is called 'scaling'.
* When we talk about scaling, this may mean running your application on multiple machines, or storing huge amounts of data, or consuming large amounts of bandwidths.

Caching

* Caching refers to storing the results of an operation so that future request return faster.
* Basically, if you have an operation that might be slow to run, say a database query, or rendering some HTML, you store the results when you run it so that you don't have to do the computation a second time. This way, you just need to reference the previous result.