Talking to Amazon.com

Ken Birman

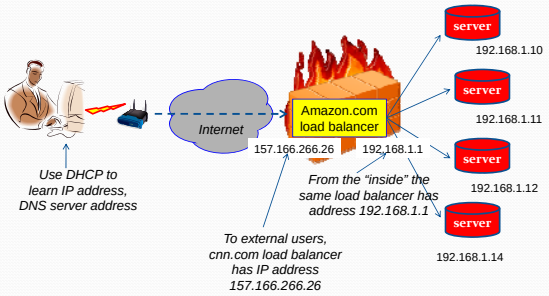
Talking to Amazon.com

- By now we understand some of the basics of talking to a big data center like Amazon.com
- Today peer a bit more deeply into the picture
  - What happens internally at Amazon.com?
  - What role is played by the "service oriented architecture" or the associated "web services standards"?
  - How does Amazon.com handle image data?
- We'll focus on Amazon when accessed via a web browser and showing you books, not some of its other lines of business (like streaming movies, hosting virtualized machines or archival storage)

Reviewing what we already know

- First, you boot your machine
  - It connects to the network, perhaps wirelessly
  - Then uses DHCP to learn its (temporary) IP address and the DNS it should talk to
  - It might also learn the address of a "web proxy"
- All of this allows it to
  - Launch a web browser
  - Connect to Amazon.com
  - Fetch a page

Reviewing what we already know



Fetching the page

- Your web browser knows how to display pages encoded in HTML, the "hypertext markup language"

```
<html>
<body>

<pre>
This is
preformatted text.
It preserves      both spaces
and line breaks.
</pre>

<p>The pre tag is good for displaying computer code:</p>

<pre>
for i = 1 to 10
    print i
next i
</pre>

</body>
</html>
```

This is
preformatted text.
It preserves both spaces
and line breaks.

The pre tag is good for displaying computer code:

```
for i = 1 to 10
    print i
next i
```


So your browser…

- Asks the DNS for the IP address of Amazon.com
  - Amazon.com itself "gives out" this address
  - Perhaps Amazon has an east and a west-coast center
  - When it first sees a request from New York, it returns the IP address of its east-coast load balancer
  - DNS will cache this and can return the same address if asked again, for a while (until the TTL expires)
- Amazon figures out that you live on the east coast from your IP address – a crude but workable approach

But Amazon is a complex system

- Years ago they discovered that no single machine could construct web pages fast enough…
  - First they expanded to have many side-by-side servers
  - But this was still too slow
  - So… they adopted an approach in which a front-end builds the page but talks to multiple back-end servers to actually obtain the content
  - Today they estimate that on average, 100 to 150 servers cooperate on *each page* that they return to a user!!!

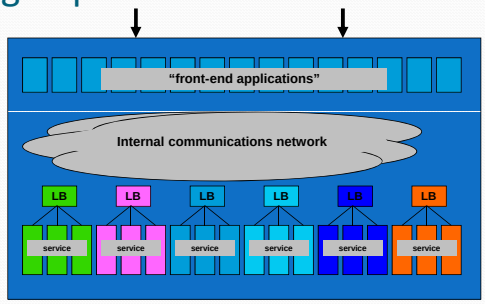
7

150 servers??? What do they do?

- One tracks down the book
- Another computes its popularity
- Another computes the price
- Another computes the inventory ("in stock")
- Another checks to see what other books people often buy when they browse this book
- Another computes your "treasure chest" of special offers

8

A glimpse inside Amazon.com



Cloud computing, web services

- Web services: the standard used to talk to the back-end services that do the real work
  - Amazon uses this between their front-end platforms (which talk HTML) and their back-end services
  - But you can also use these web services directly from your client computer and talk directly to many of those services
  - Amazon is promoting this as a way that end-users can build Amazon-hosted applications and platforms
  - A rapidly growing secondary market of developers who are extending Amazon's reach
- The broad term for this is "cloud computing"

10

Cloud computing

- Wikipedia:

Cloud computing is Internet ("cloud") based development and use of computer technology ("computing").

It is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet.

Users need not have knowledge of, expertise in, or control over the technology infrastructure "in the cloud" that supports them.

11

Web Services

- Wikipedia:

A Web service (also Web Service) is defined by the W3C as "a software system designed to support interoperable machine-to-machine interaction over a network".

Web services are frequently just Web APIs that can be accessed over a network, such as the Internet, and executed on a remote system hosting the requested services.

12

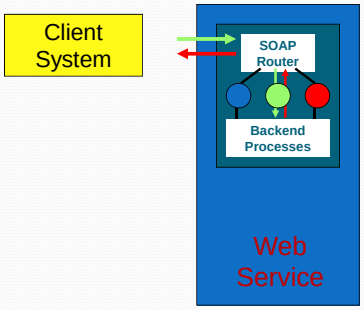
Service Oriented Architectures

In computing, service-oriented architecture (SOA) provides methods for systems development and integration where systems group functionality around business processes and package these as interoperable services.

A SOA infrastructure allows different applications to exchange data with one another. This allows a variety of applications to be constructed using a shared set of reusable components.

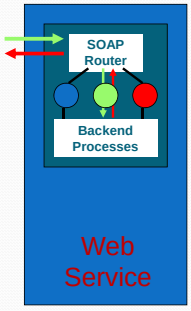
13

Basic Web Services model



Basic Web Services model

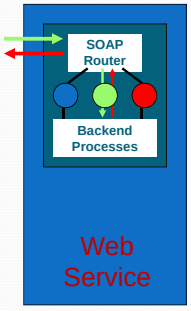
- "Web Services are software components described via WSDL which are capable of being accessed via standard network protocols such as SOAP over HTTP."



Basic Web Services model

- "Web Services are software components described via WSDL which are capable of being accessed via standard network protocols such as SOAP over HTTP."

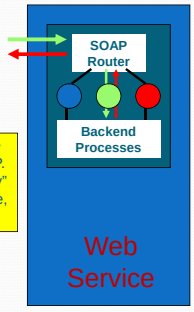
Today, SOAP is the primary standard. SOAP provides rules for encoding the request and its arguments.



Basic Web Services model

- "Web Services are software components described via WSDL which are capable of being accessed via standard network protocols such as SOAP over HTTP."

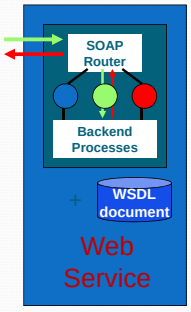
Similarly, the architecture doesn't assume that all access will employ HTTP over TCP. In fact, .NET uses Web Services "internally" even on a single machine. But in that case, communication is over COM



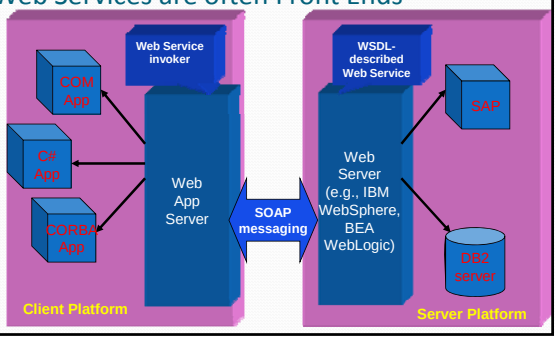
Basic Web Services model

- "Web Services are software components described via WSDL which are capable of being accessed via standard network protocols such as SOAP over HTTP."

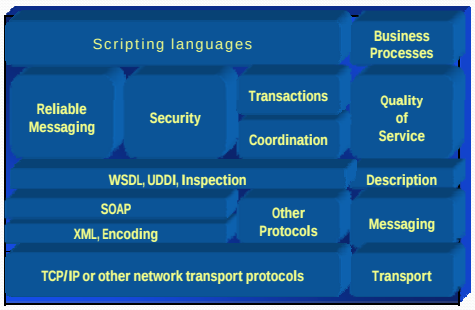
WSDL documents are used to drive object assembly, code generation, and other development tools.



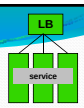
Web Services are often Front Ends



The Web Services "stack"



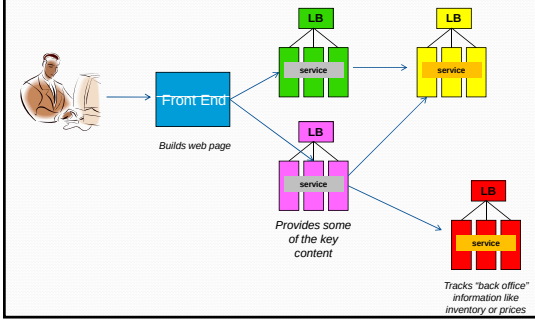
More complications



- How does Amazon build scalable back-end services?
  - They develop their applications to run in a "clustered" manner with multiple server instances
  - The platform varies the number depending on load
- A load balancer spreads the work
- Each service may in turn talk to other services, make use of data stored in files or databases, etc
  - So you should think of a graph of services

21

Example: A graph of services



22

What about images?

- Handling of images, videos is "special", and same for advertising content
  - Many companies prefer to outsource the management of this kind of content
  - For example cnn.com would rather not keep all the photos on their own web site
- How do they do it?

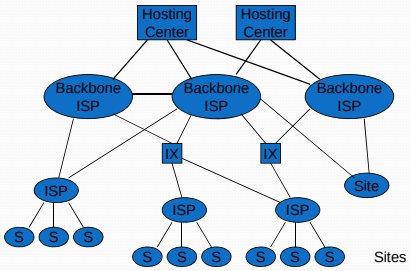
23

Content hosting services

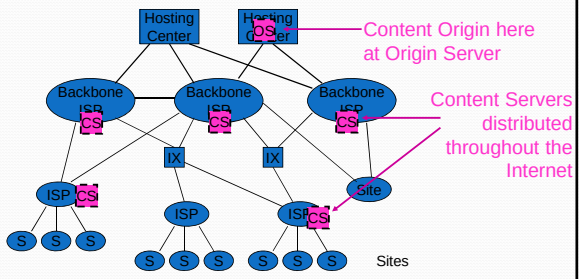
- There are some companies that specialize in "hosting" images and similar content
  - Photos and other large pictures
  - Advertisements
  - Videos (even entire episodes of Fringe or Desparate Housewives….)
- These companies often run large numbers of small data centers at many locations world wide
  - Typical example: Akamai.com

24

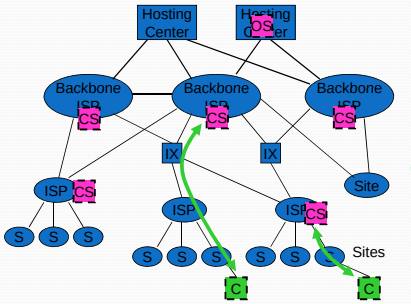
Content Routing Principle
(a.k.a. Content Distribution Network)



Content Routing Principle
(a.k.a. Content Distribution Network)



Content Routing Principle
(a.k.a. Content Distribution Network)



How it works

- Instead of including images in the web page sent to your browser, cnn.com (or whoever) includes URLs that tell the browser where to fetch the images
- The browser downloads the page… then as it renders it, fetches these images
  - It does this in parallel, so it may end up with 30 or 50 parallel transfers underway
- These URLs point to the image but within Akamai.com, not cnn.com

28

Akamai.com

- Akamai uses various tricks to "redirect" the request to a server in its network
  - Ideally, one close to you (so download will be fast)
  - And not too heavily loaded
  - If needed, their server can fetch a copy of the content on demand. Then it saves that copy for future reuse
- Akamai may have millions of machines playing this role at any point of time!
  - Each can simultaneously send images to perhaps 50 users, so they can handle tens of millions of simultaneous downloads
  - Akamai is just one of many companies that do this

29

So: You access cnn.com….

- But your data comes back from many places
  - Cnn.com itself
    - Within it, perhaps assembled from many servers
  - Akamai.com
  - Doubleclick.com – advertising placement and tracking
- Advertising: often inserted by specialists that try and place appropriate advertising based on profiles of you
  - Biking stuff for me, spring break tee shirts for someone else, investment suggestions for yet another person
  - Rewarded if you click that ad!

30

Cookies

- Many web platforms leave small files on your computer as notes "to themselves"
  - These are called cookies
  - Uses to remember that you've visited cnn.com before, logged in as KenBirman, password Biscuit, focused on the science web pages, etc
  - Like a mini user profile
- When your browser connects, it automatically sends the cookie contents as part of the new session protocol

31

Cookie: Example

```
Set-Cookie: RMID=732423sdfs73242; expires=Fri, 31-Dec-2010 23:59:59 GMT; path=/; domain=.example.net
```

- The name of this cookie is RMID
- Its value is the string 732423sdfs73242.
  - The server can use an arbitrary string as the cookie value
  - It can collapse multiple variables in a single string:

    a=12&b=abcd&c=32
- The path (/) and domain (.example.net) tell the browser to send this cookie on every page request to any server in domain "example.net"

32

Cookies

- Used to track
  - Who you are ("Welcome back, Ken!")
  - What you've done in the past ("Still interested in cameras?")
  - When you last visited
- But keep in mind that sites may have other ways to track you too, even if cookies are disabled
  - IP address (not reliable but still a good hint)
  - May just insist that you log in

33

Cookies

- Cookies can contain things you think of as private
- So cookie is associated with a specific site.
  - Just the same, when talking to a site over HTTP, anyone spying on the network can see the cookie pass by in plain ASCII text
  - For secure sites (HTTPS), other sites shouldn't be able to "spy" on cookies they don't own (unless brower is buggy)
- Cookie offers a quick way to "look up" the user so that site can personalize the browsing experience

34

Recap

- You thought you were talking to Amazon.com, or cnn.com
  - Actually, you talked to one of their many data centers
  - Within that center, to a collection of machines that may have included *hundreds* of mini-services
  - All of this resulted in the web page your browser rendered… but that in turn may have left image content to be fetched from a content hosting service like Akamai
- Effect? Massive parallelism. Hundreds of machines cooperating to render that one page!

35

Javascript/AJAX

- Used to implement the famous Gieco chameleon…
  - Javascript is a programming language, unrelated to Java
  - AJAX is a kind of portable operating system: **A**synchronous **Ja**va for Remote E**x**ecution
- People use it to create animated images and other fancy content
  - Google Earth uses Javascript to do pan/zoom/selectable layers for their downloads
  - Geico uses it to implement the dancing lizard
  - Increasingly common to send very sophisticated programs to your web browser

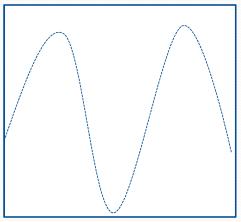
36

Javascript/AJAX Example

```
<table>
  <tr><td>Change value for view result</td></tr>
  <tr><td><form><input value="100"
onchange="tg.onchange(this.value)"><input type=button
value="Change"></form> [-100..100]</td></tr>
</table>
<script type=text/javascript language=javascript>


function Scatter() {

  this.range = [0,1];
  this.top = 0;
  this.id = "myChart";
  this.left = 0;
  this.height = 30;
  this.width = 400;
  this.borderWidth = 2;
  this.borderStyle = "outset";
  this.linewidth = 1;
  this.parent = null;
  this.hilightColor = "navy";

  this.k = 100;

  this.onchange = function (newValue) {
   newValue = parseInt(newValue);
   if(newValue>=-100 && newValue<=100) {
     this.k = newValue;
     this.redraw();
   }
  }
  this.getWrapperHTML = function () {
   with(this)
    return "<div style='position:absolute;left:" + left + "px;" +
                    "top:" + top + "px;" +
                    "width:" + width + "px;" +
                    "height:" + height + "px;" +
                    "border-style:" + borderStyle + ";" +
                    "border-width:" + borderWidth + "px;'" +
                    " id=" + id + "></div>";
  }


  this.values = [[0,0]];
  this.redraw = function() {
   var tempstr = "";
   values = [];
   for(var i=0;i<200;i++) {
    x = i;
    y = 150 + k * Math.sin (i/30);
    values[values.length] = [x, y];
   }


   for(var i=0; i<values.length; i++) {
     tempstr += "<div style='position:absolute;background-Color:" +
hilightColor +
       ";left:" + (borderWidth + parseInt(values[i][0])) + "px;" +
       "top:" + (height - 2 * borderWidth - parseInt(values[i][1])) + "px;" +
       "width:" + linewidth + "px;height:" + linewidth + "px;font-
size:0px'></div>";
   }
   document.getElementById(this.id).innerHTML = tempstr;
  }
 }

 this.create = function() {
  document.body.innerHTML += this.getWrapperHTML();
  this.redraw();
 }
}

var tg;
function delay_this(){
tg = new Scatter();
with(tg) {
 top = 70;
 left = 15;
 width = 300;
 height = 300;
 create();
}}
setTimeout("delay_this()",3000);
/*
 * It is possible to remove this delay call call the delay_this() routine
 * directly here
 */
delay_this();
*/

</script>
```


Javascript/AJAX Example

- Example draws a little sine-wave graph
- In general, Javascript can implement programmed effects and behaviors
- Can also access cookies and even files, depending on how you set permissions
- Some people consider it to be a true "distribued O/S"!



Additional complications

- Some web pages are modified "on the fly"
  - For example, in the network itself
  - Google, ISPs all want to do this… they want to insert hyperlinks that you can click (and that they can use to show advertising)
- Effect? The web page you download might not be identical to what the web site sent you!

Things to think about

- None of this is very secure
  - This is why we switch to https for transactions
  - It uses encryption on the browser/server connection
- But with Javascript there are more and more security loopholes and complications
  - Basically, the sophistication of the options is way beyond what we understand how to protect
  - This is in the nature of technology: features are more rewarded than robustness, security

Summary

- Modern web browser is a new kind of operating system!
  - A network operating system
  - Programs are "loaded" over the network, then execute inside browser windows
- More and more of what we do involves browser-accessed applications
  - So-called Cloud Computing
- So this new kind of O/S needs our attention…