

- [About Us](#)
- [About You](#)
- [Our Contributor Team](#)
- [Purpose Index](#)

- Exclusive updates on:



Facilitating the spread of knowledge and innovation in enterprise software development

[Login](#)



- [En](#) |
- [中文](#) |
- [日本](#) |
- [Fr](#) |
- [Br](#)

767,446 May unique visitors

- [Development](#)
 - [Java](#)
 - [.Net](#)
 - [Cloud](#)
 - [Mobile](#)
 - [HTML5](#)
 - [JavaScript](#)
 - [Ruby](#)
 - [DSLs](#)
 - [Python](#)
 - [PHP](#)
 - [API](#)

Featured in Development

[Building Hypermedia APIs with HTML](#)



Jon Moore discusses the pros and cons using HTML for an API media type, as well as the particulars of implementation patterns, from representations to building programmatic HTML clients.

[All in Development](#)

- [Architecture & Design](#)
 - [Modeling](#)
 - [Scalability/Performance](#)
 - [DDD](#)
 - [BDD](#)
 - [AOP](#)
 - [Patterns](#)
 - [Security](#)
 - [Cloud](#)
 - [SOA](#)

Featured in Architecture & Design

[Jepsen: Testing the Partition Tolerance of PostgreSQL, Redis, MongoDB and Riak](#)



Distributed systems are characterized by exchanging state over high-latency or

unreliable links. The system must be robust to both node and network failure if it is to operate reliably--however, not all systems satisfy the safety invariants we'd like. In this article, we'll explore some of the design considerations of distributed databases, and how they respond to network partitions.

[All in Architecture & Design](#)

- [Process & Practices](#)
 - [Agile](#)
 - [Leadership](#)
 - [Collaboration](#)
 - [Agile Techniques](#)
 - [Methodologies](#)
 - [Continuous Integration](#)
 - [Lean/Kanban](#)

Featured in Process & Practices

[Spock: A Highly Logical Way To Test](#)



Howard Lewis Ship overviews Spock, a testing framework for the JVM.

[All in Process & Practices](#)

- [Operations & Infrastructure](#)
 - [Performance](#)
 - [Big Data](#)
 - [DevOps](#)
 - [Cloud](#)
 - [APM](#)
 - [Virtualization](#)
 - [NoSQL](#)

Featured in Operations & Infrastructure

[Jepsen: Testing the Partition Tolerance of PostgreSQL, Redis, MongoDB and Riak](#)



Distributed systems are characterized by exchanging state over high-latency or unreliable links. The system must be robust to both node and network failure if it is to operate reliably--however, not all systems satisfy the safety invariants we'd like. In this article, we'll explore some of the design considerations of distributed databases, and how they respond to network partitions.

[All in Operations & Infrastructure](#)

- [Enterprise Architecture](#)
 - [Enterprise Architecture](#)
 - [BPM](#)
 - [Business/IT Alignment](#)
 - [Architecture Documentation](#)
 - [IT Governance](#)
 - [SOA](#)

Featured in Enterprise Architecture

[Java Garbage Collection Distilled](#)



CMS, G1, Young Gen, New Gen, Old Gen, Eden, and the hundreds of JVM start-up flags... does this all baffle you when trying to tune the garbage collector to get the required throughput and latency from your Java application? Don't worry, you are not alone. This article will attempt to explain the tradeoffs when choosing and tuning garbage collection algorithms for a particular workload.

[All in Enterprise Architecture](#)

[New York 2013](#)

[Jun 10th - Jun 14th](#)

- [Mobile](#)
- [HTML5](#)
- [JavaScript](#)
- [APM](#)
- [Cloud](#)
- [NoSQL](#)
- [SOA](#)
- [DevOps](#)

[All topics](#)

[New UI](#)

You are here: [InfoQ Homepage](#) [Articles](#) How to GET a Cup of Coffee

How to GET a Cup of Coffee

Posted by [Jim Webber, Savas Parastatidis & Ian Robinson](#) on Oct 02, 2008 | [16 comments](#)

- Share



- |
-
-
-
-

-
-
-
-
- ["Read later"](#)
- ["My Reading List"](#)

We are used to building distributed systems on top of large middleware platforms like those implementing CORBA, the Web Services protocols stack, J2EE, etc. In this article, we take a different approach, treating the protocols and document formats that make the Web tick as an application platform, which can be accessed through lightweight middleware. We showcase the role of the Web in application integration scenarios through a simple customer-service interaction scenario. In this article, we use the Web as our primary design philosophy to distil and share some of the thinking in our forthcoming book “GET /connected - Web-based integration” (working title).

Introduction

The integration domain as we know it is changing. The influence of the Web and the trend towards more agile practices are challenging our notions of what constitutes good integration. Instead of being a specialist activity conducted in the void between systems – or even worse, an afterthought – integration is now an everyday part of successful solutions.

Yet, the impact of the Web is still widely misunderstood and underestimated in enterprise computing. Even those who are Web-savvy often struggle to understand that the Web isn't about middleware solutions supporting XML over HTTP, nor is it a crude RPC mechanism. This is a shame because the Web has much more value than simple point-to-point connectivity; it is in fact a robust integration platform.

Related Vendor Content

[Why HTML5 is Not Enough](#)

[A Unified Mobile Architecture for the Modern Data Center](#)

[Enterprise-class API Patterns for Cloud & Mobile](#)

[451 Group Analyst Report: Review of Intel & Mashery API Management Platform](#)

[API Management - Application Security Made Easy](#)

Related Sponsor

*[Cloud API Management – the Service Gateway Approach](#)

[*451 Group Analyst Report: Review of Intel & Mashery API Management Platform](#)



In this article we'll showcase some interesting uses of the Web, treating it as a pliant and robust platform for doing very cool things with enterprise systems. And there is nothing that typifies enterprise software more than workflows...

Why Workflows?

Workflows are a staple of enterprise computing, and have been implemented in middleware practically forever (at least in computing terms). A workflow structures work into a number of discrete steps and the events that prompt transitions between steps. The overarching business process implemented by a workflow often spans several enterprise information systems, making workflows fertile ground for integration work.

Starbucks: Standard generic coffee deserves standard generic integration

If the Web is to be a viable technology for enterprise (and wider) integration, it has to be able to support workflows – to reliably coordinate the interactions between disparate systems to implement some larger business capability.

To do justice to a real-world workflow, we'd no doubt have to address a wealth of technical and domain-specific details, which would likely obscure the aim of this article, so we've chosen a more accessible domain to illustrate how Web-based integration works: Gregor Hohpe's Starbucks coffee shop workflow. [In his popular blog posting](#), Gregor describes how Starbucks functions as a decoupled revenue-generating pipeline:

"Starbucks, like most other businesses is primarily interested in maximizing throughput of orders. More orders equals more revenue. As a result they use asynchronous processing. When you place your order the cashier marks a coffee cup with your order and places it into the queue. The queue is quite literally a queue of coffee cups lined up on top of the espresso machine. This queue decouples cashier and barista and allows the cashier to keep taking orders even if the barista is backed up for a moment. It allows them to deploy multiple baristas in a Competing Consumer scenario if the store gets busy."

While Gregor prefers EAI techniques like message-oriented middleware to model Starbucks, we'll model the same scenario using Web resources – addressable entities that support a uniform interface. In fact, we'll show how Web techniques can be used with all the dependability associated with traditional EAI tools, and how the Web is much more than XML messaging over a request/response protocol!

We'll apologise in advance for taking liberties with the way Starbucks works because our goal here isn't to model Starbucks completely accurately, but to illustrate workflows with Web-based services. So with belief duly suspended, let's jump in.

Stating the Obvious

Since we're talking about workflows, it makes sense to understand the states from which our workflows are composed, together with the events that transition the workflows from state to state. In our example, there are two workflows, which we've modelled as state machines. These workflows run concurrently. One models the interaction between the customer and the Starbucks service as shown in Figure 1 the other captures the set of actions performed by a barista as per Figure 2.

In the customer workflow, customers advance towards the goal of drinking some coffee by interacting with the Starbucks service. As part of the workflow, we assume that the customer places an order, pays, and then waits for their drink. Between placing and paying for the order, the customer can usually amend it – by, for example, asking for semi-skimmed milk to be used.

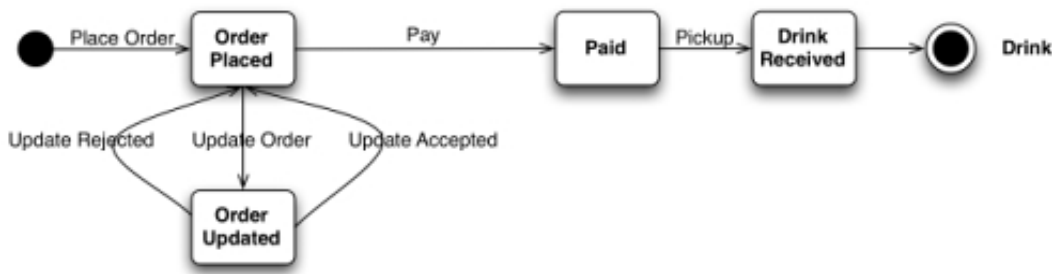


Figure1 The Customer State Machine

The barista has his or her own state machine, though it's not visible to the customer; it's private to the service's implementation. As shown in Figure 2, the barista loops around looking for the next order to be made, preparing the drink, and taking the payment. An instance of the loop can begin when an order is added to the barista's queue. The outputs of the workflow are available to the customer when the barista finishes the order and releases the drink.



Figure 2 The Barista's State Machine

Although all of this might seem a million miles away from Web-based integration, each transition in our two state machines represents an interaction with a Web resource. Each transition is the combination of a HTTP verb on a resource via its URI causing state changes.

GET and HEAD are special cases since they don't cause state transitions. Instead they allow us to inspect the current state of a resource.

But we're getting ahead of ourselves. Thinking about state machines and the Web isn't easy to swallow in one big lump. So let's revisit the entire scenario from the beginning, look at it in a Web context, and proceed one step at a time.

The Customer's Viewpoint

We'll begin at the beginning, with a simple story card that kick-starts the whole process:



This story contains a number of useful actors and entities. Firstly, there's the customer actor, who is the obvious consumer of the (implicit) Starbucks service. Secondly, there are two interesting entities (coffee and order), and an interesting interaction (ordering), which starts our workflow.

To submit an order to Starbucks, we simply POST a representation of an order to the well-known Starbucks ordering URI, which for our purposes will be `http://starbucks.example.org/order`.

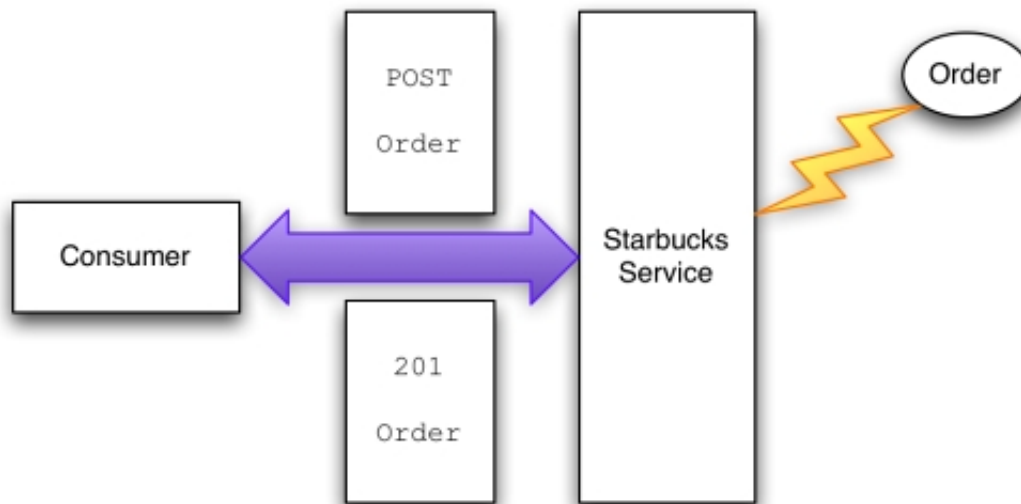


Figure 3 Ordering a coffee

Figure 3 shows the interaction to place an order with Starbucks. Starbucks uses an XML dialect to represent entities from its domain; interestingly, this dialect also allows information to be embedded so that customers can progress through the ordering process – as we'll see shortly. On the wire the act of posting looks something like Figure 4.

*In the human Web, consumers and services use HTML as a representation format. HTML has its own particular semantics, which are understood and adopted by all browsers: `<a/>`, for example, **means** “an anchor that links to another document or to a bookmark within the same document.” The consumer application – the Web browser – simply renders the HTML, and the state machine (that's you!) follows links using *GET* and *POST*. In Web-based integration the same occurs, except the services and their consumers not only have to agree on the interaction protocols, but also on the format and semantics of the representations.*

```
POST /order HTTP 1.1
Host: starbucks.example.org
Content-Type: application/xml
Content-Length: ...

<order xmlns="http://starbucks.example.org/">
  <drink>latte</drink>
</order>
```

Figure 4 POSTing a drinks order

The Starbucks service creates an order resource, and then responds to the consumer with the location of this new resource in the Location HTTP header. For convenience, the service also places the representation of the newly created order resource in the response. The response looks something like .

```
201 Created
Location: http://starbucks.example.org/order/1234
Content-Type: application/xml
Content-Length: ...

<order xmlns="http://starbucks.example.org/">
  <drink>latte</drink>
  <cost>3.00</cost>
  <next xmlns="http://example.org/state-machine"
    rel="http://starbucks.example.org/payment"
    uri="https://starbucks.example.com/payment/order/1234"
    type="application/xml"/>
</order>
```

Figure 5 Order created, awaiting payment

The 201 Created status indicates that Starbucks successfully accepted the order. The Location header gives the URI of the newly created order. The representation in the response body contains confirmation of what was ordered along with the cost. In addition, this representation contains the URI of a resource with which Starbucks expects us to interact to make forward progress with the customer workflow; we'll use this URI later.

Note that the URI is contained in a `<next/>` tag, not an HTML `<a/>` tag. `<next/>` is here meaningful in the context of the customer workflow, the semantics of which have been agreed a priori.

We've already seen that the 201 Created status code indicates the successful creation of a

resource. We'll need a handful of other useful codes both for this example and for Web-based integration in general:

200 OK - This is what we like to see: everything's fine; let's keep going. *201 Created* - We've just created a resource and everything's fine.

202 Accepted - The service has accepted our request, and invites us to poll a URI in the Location header for the response. Great for asynchronous processing.

303 See Other - We need to interact with a different resource. We're probably still OK.

400 Bad Request - We need to reformat the request and resubmit it.

404 Not Found - The service is far too lazy (or secure) to give us a real reason why our request failed, but whatever the reason, we need to deal with it.

409 Conflict - We tried to update the state of a resource, but the service isn't happy about it. We'll need to get the current state of the resource (either by checking the response entity body, or doing a GET) and figure out where to go from there.

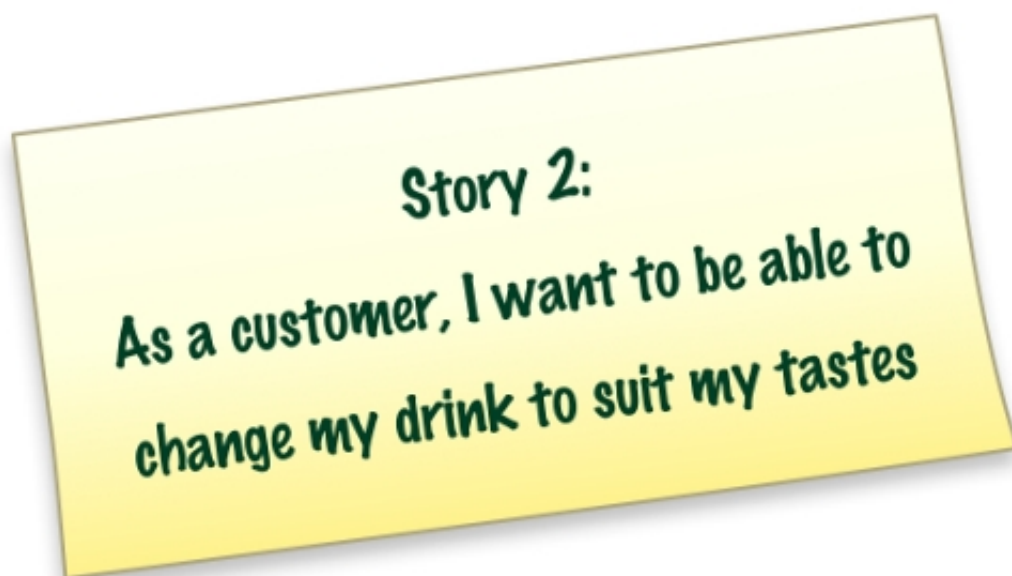
412 Precondition Failed - The request wasn't processed because an Etag, If-Match or similar guard header failed evaluation. We need to figure out how to make forward progress.

417 Expectation Failed - You did the right thing by checking, but please don't try to send that request for real.

500 Internal Server Error - The ultimate lazy response. The server's gone wrong and it's not telling why. Cross your fingers...

Updating an Order

One of the nice things about Starbucks is you can customise your drink in a myriad of different ways. In fact, some of the more advanced customers would be better off ordering by chemical formula, given the number of upgrades they demand! But let's not be that ambitious – at least not to start with. Instead, we'll look at another story card:



Looking back on Figure 4, it's clear we made a significant error: for anyone that really likes coffee, a single shot of espresso is going to be swamped by gallons of hot milk. We're going to have to change that. Fortunately, the Web (or more precisely HTTP) provides support for such changes, and so does our

service.

Firstly, we'll make sure we're still allowed to change our order. Sometimes the barista will be so fast our coffee's been made before we've had a chance to change it – and then we're stuck with a cup of hot coffee-flavoured milk. But sometimes the barista's a little slower, which gives us the opportunity to change the order before the barista processes it. To find out if we can change the order, we ask the resource what operations it's prepared to process using the HTTP OPTIONS verb, as shown on the wire in Figure 6.

Request	Response
OPTIONS /order/1234 HTTP 1.1 Host: starbucks.example.org	200 OK Allow: GET, PUT

Figure6 Asking for OPTIONS

From Figure 6 we see that the resource is readable (it supports GET) and it's updatable (it supports PUT). As we're good citizens of the Web, we can, optionally, do a trial PUT of our new representation, testing the water using the Expect header before we do a real PUT – like in Figure 7.

Request	Response
PUT /order/1234 HTTP 1.1 Host: starbucks.example.com Expect: 100-Continue	100 Continue

Figure 7 Look before you leap!

If it had no longer been possible to change our order, the response to our “look before you leap” request in Figure 7 would have been 417 Expectation Failed. But here the response is 100 Continue, which allows us to try to PUT an update to the resource with an additional shot of espresso, as shown in Figure 8. PUTting an updated resource representation effectively changes the existing one. In this instance PUT lodges a new description with an <additions/> element containing that vital extra shot.

Although partial updates are the subject of deep philosophical debates within the REST community, we take a pragmatic approach here and assume that our request for an additional shot is processed in the context of the existing resource state. As such there is little point in moving the whole resource representation across the network for each operation and so we transmit deltas only.

```
PUT /order/1234 HTTP 1.1
Host: starbucks.example.com
Content-Type: application/xml
Content-Length: ...

<order xmlns="http://starbucks.example.org/">
  <additions>shot</additions>
</order>
```

Figure 8 Updating a resource's state

If we're successfully able to PUT an update to the new resource state, we get a 200 response from the server,

as shown in Figure 9.

```
200 OK
Location: http://starbucks.example.com/order/1234
Content-Type: application/xml
Content-Length: ...

<order xmlns="http://starbucks.example.org/">
  <drink>latte</drink>
  <additions>shot</additions>
  <cost>4.00</cost>
  <next xmlns="http://example.org/state-machine"
    rel="http://starbucks.example.org/payment"
    uri="https://starbucks.example.com/payment/order/1234"
    type="application/xml"/>
</order>
```

Figure 9 Successfully updating the state of a resource

Checking `OPTIONS` and using the `Expect` header can't totally shield us from a situation where a change at the service causes subsequent requests to fail. As such we don't mandate their use, and as good Web citizens we're going to handle 405 and 409 responses anyway.

`OPTIONS` and especially using the `Expect` header should be considered optional steps.

Even with our judicious use of `Expect` and `OPTIONS`, sometimes our `PUT` will fail; after all, we're in a race with the barista – and sometimes those guys just fly!

If we lose the race to get our extra shot, we'll learn about it when we try to `PUT` the updates to the resource. The response in Figure 10 is typical of what we can expect. `409 Conflict` indicates the resource is in an inconsistent state to receive the update. The response body shows the difference between the representation we tried to `PUT` and the resource state on the server side. In coffee terms it's too late to add the shot – the barista's already pouring the hot milk.

```
409 Conflict

<order xmlns="http://starbucks.example.org/">
  <drink>latte</drink>
  <cost>3.00</cost>
  <next xmlns="http://example.org/state-machine"
    rel="http://starbucks.example.org/payment"
    uri="https://starbucks.example.com/payment/order/1234"
    type="application/xml"/>
</order>
```

Figure 10 Losing a race

We've discussed using `Expect` and `OPTIONS` to guard against race conditions as much as possible. Besides these, we can also attach `If-Unmodified-Since` or `If-Match` headers to our `PUT` to convey our intentions to the receiving service. `If-Unmodified-Since` uses the timestamp and `If-Match` the `ETag`¹ of the original order. If the order hasn't changed since we created it – that is, the barista hasn't started preparing our coffee yet – then the change will be processed. If the order has changed, we'll get a 412

Precondition Failed response. If we lose the race, we're stuck with milky coffee, but at least we've not transitioned the resource to an inconsistent state.

There are a number of patterns for consistent state updates using the Web. HTTP PUT is idempotent, which takes much of the intricate work out of updating state, but there are still choices that need to be made. Here's our recipe for getting updates right:

- 1. Ask the service if it's still possible to PUT by sending OPTIONS. This step is optional. It gives clients a clue about which verbs the server supports for the resource at the time of asking, but there are no guarantees the service will support those same verbs indefinitely.*
- 2. Use an If-Unmodified-Since or If-Match header to help the server guard against executing an unnecessary PUT. You'll get a 412 Precondition Failed if the PUT subsequently fails. This approach depends either on slowly changing resources (1 second granularity) for If-Unmodified-Since or support for ETags for If-Match.*
- 3. Immediately PUT the update and deal with any 409 Conflict responses. Even if we use (1) and (2), we may have to deal with these responses, since our guards and checks are optimistic in nature.*

The W3C has [a non-normative note](#) on detecting and dealing with inconsistent updates that argues for using ETag. ETags are our preferred approach.

After all that hard work updating our coffee order, it seems only fair that we get our extra shot. So for now let's go with our happy path, and assume we managed to get our additional shot of espresso. Of course, Starbucks won't hand our coffee over unless we pay (and it turns out they've already hinted as much!), so we need another story:



Remember the `<next/>` element in the response to our original order? This is where Starbucks embedded

information about another resource in the order representation. We saw the tag earlier, but chose to ignore it while correcting our order. But now it's time to look more closely at it:

```
<next xmlns="http://example.org/state-machine"
  rel="http://starbucks.example.org/payment"
  uri="https://starbucks.example.com/payment/order/1234"
  type="application/xml"/>
```

There are a few aspects to the next element worth pointing out. First is that it's in a different namespace because state transitions are not limited to Starbucks. In this case we've decided that such transition URIs should be held in a communal namespace to facilitate re-use (or even eventual standardisation).

Then, there's the embedded semantic information (a private microformat, if you like) in the `rel` attribute. Consumers that understand the semantics of the `http://starbucks.example.org/payment` string can use the resource identified by the `uri` attribute to transition to the next state (payment) in the workflow.

The `uri` in the `<next/>` element points to a payment resource. From the `type` attribute, we already know the expected resource representation is XML. We can work out what to do with the payment resource by asking the server which verbs that resource supports using `OPTIONS`.

Microformats are a way to embed structured, semantically-rich data inside existing documents. Microformats are most common in the human readable Web, where they are used to add structured representations of information like calendar events to Web pages. However, they can just as readily be turned to integration purposes. Microformat terminology is agreed by the microformats community, but we are at liberty to create our own private microformats for domain-specific semantic markup.

Innocuous as they seem, simple links like the one of Figure 10 are the crux of what the REST community rather verbosely calls “Hypermedia as the engine of application state.” More simply, URIs represent the transitions within a state machine. Clients operate application state machines, like the ones we saw at the beginning of this article, by following links.

Don't be surprised if that takes a little while to sink in. One of the most surprising things about this model is the way state machines and workflows gradually describe themselves as you navigate through them, rather than being described upfront through WS-BPEL or WS-CDL. But once your brain has stopped somersaulting, you'll see that following links to resources allows us to make forward progress in our application's various states. At each state transition the current resource representation includes links to the next set of possible resources and the states they represent. And because those next resources are just Web resources, we already know what to do with them.

Our next step in the customer workflow is to pay for our coffee. We know the total cost from the `<cost/>` element in the order, but before we send payment to Starbucks we'll ask the payment resource how we're meant to interact with it, as shown in Figure 11.

How much upfront knowledge of a service does a consumer need? We've already suggested that services and consumers will need to agree the semantics of the representations they exchange prior to interacting. Think of these representation formats as a set of possible states and transitions. As a consumer interacts with a service, the service chooses states and

transitions from the available set and builds the next representation. The process – the “how” of getting to a goal – is discovered on the fly; what gets wired together as part of that process is, however, agreed upfront.

Consumers typically agree the semantics of representations and transitions with a service during design and development. But there's no guarantee that as service evolves, it won't confront the client with state representations and transitions the client had never anticipated but knows how to process – that's the nature of the loosely coupled Web. Reaching agreement on resource formats and representations under these circumstances is, however, outside the scope of this article.

Our next step is to pay for our coffee. We know the total cost of our order from the <cost> element embedded in the order representation, and so our next step is to send a payment to Starbucks so the barista will hand over the drink. Firstly we'll ask the payment resource how we're meant to interact with it, as shown in Figure 11.

Request	Response
OPTIONS/payment/order/1234 HTTP 1.1 Host: starbucks.example.com	Allow: GET, PUT

Figure 11 Figuring out how to pay

The response indicates we can either read (via GET) the payment or update it (via PUT). Knowing the cost, we'll go ahead and PUT our payment to the resource identified by the payment link. Of course, payments are privileged information, so we'll protect access to the resource by requiring authentication².

Request
<pre>PUT /payment/order/1234 HTTP 1.1 Host: starbucks.example.com Content-Type: application/xml Content-Length: ... Authorization: Digest username="Jane Doe" realm="starbucks.example.org" nonce="..." uri="payment/order/1234" qop=auth nc=00000001 cnonce="..." reponse="..." opaque="..." <payment xmlns="http://starbucks.example.org/"> <cardNo>123456789</cardNo> <expires>07/07</expires> <name>John Citizen</name> <amount>4.00</amount> </payment></pre>
Response


```
201 Created
Location: https://starbucks.example.com/payment/order/1234
Content-Type: application/xml
Content-Length: ...

<payment xmlns="http://starbucks.example.org/">
  <cardNo>123456789</cardNo>
  <expires>07/07</expires>
  <name>John Citizen</name>
  <amount>4.00</amount>
</payment>
```

Figure 12 Paying the bill

For successful payments, the exchange shown in Figure 12 is all we need. Once the authenticated PUT has returned a 201 Created response, we can be happy the payment has succeeded, and can move on to pick up our drink.

But things can go wrong, and when money is at stake we'd rather things either didn't go wrong or are recoverable when they do³. A number of things can obviously go wrong with our payment:

- We can't connect to the server because it is down or unreachable;
- The connection to the server is severed at some point during the interaction;
- The server returns an error status in the 4xx or 5xx range.

Fortunately, the Web helps us in each of these scenarios. For the first two cases (assuming the connectivity issue is transient), we simply PUT the payment again until we receive a successful response. We can expect a 200 response if a prior PUT had in fact succeeded (effectively an acknowledgement of a no-op from the server) or a 201 if the new PUT eventually succeeds in lodging the payment. The same holds true in the third case where the server responds with a 500, 503 or 504 response code.

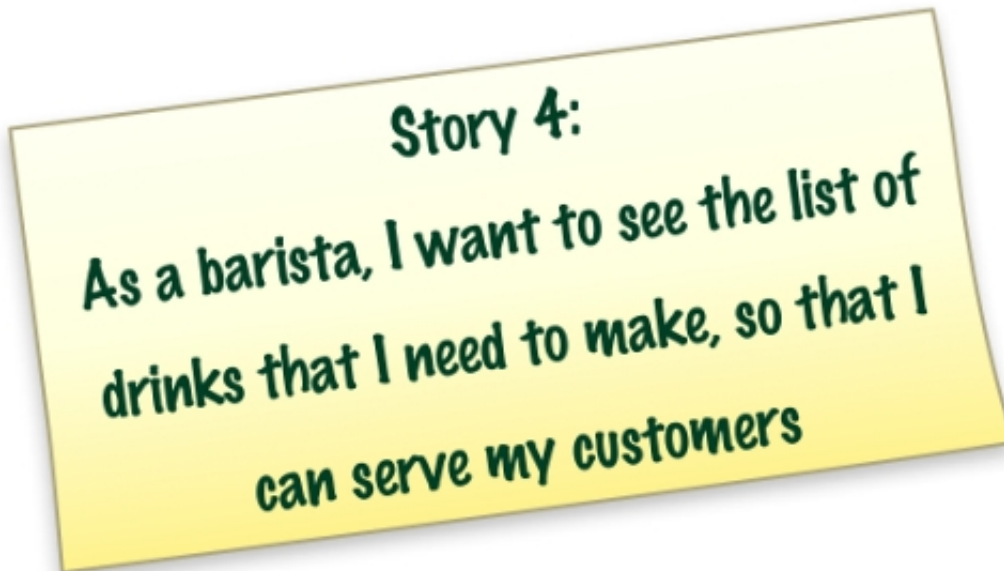
Status codes in the 4xx range are trickier, but they still indicate how to make forward progress. For example, a 400 response indicates that we PUT something the server doesn't understand, and should rectify our payload before PUTting it again. Conversely, a 403 response indicates that the server understood our request but is refusing to fulfil it and doesn't want us to re-try. In such cases we'll have to look for other state transitions (links) in the response payload to make alternative forward progress.

We've used status codes several times in this example to guide the client towards its next interaction with the service. Status codes are semantically rich acknowledgments. By implementing services that produce meaningful status codes and clients that know how to handle them, we can layer a coordination protocol on top of HTTP's simple request-response mechanism, adding a high degree of robustness and reliability to distributed systems.

Once we've paid for our drink we've reached the end of our workflow, and the end of the story as far as the consumer goes. But it's not the end of the whole story. Let's now go inside the service boundary, and look at Starbucks' internal implementation.

The Barista's Viewpoint

As customers we tend to put ourselves at the centre of the coffee universe, but we're not the only consumers of a coffee service. We know already from our “race” with the barista that the service serves at least one other set of interested parties, not the least of which is the barista. In keeping with our incremental delivery style, it's time for another story card.



Lists of drinks are easily modelled using Web formats and protocols. Atom feeds are a perfectly good format for lists of practically anything, including outstanding coffee orders, so we'll adopt them here. The barista can access the Atom feed with a simple GET on the feed's URI, which for outstanding orders is <http://starbucks.example.org/orders> in Figure 13.

```
200 OK
Expires: Thu, 12Jun2008 17:20:33 GMT
Content-Type: application/atom+xml
Content-Length: ...

<?xml version="1.0" ?>
<feed xmlns="http://www.w3.org/2005/Atom">
  <title>Coffees to make</title>
  <link rel="alternate"
        uri="http://starbucks.example.org/orders"/>
  <updated>2008-06-10T19:18:43Z</updated>
  <author><name>Barista System</name></author>
  <id>urn:starkbucks:barista:coffees-to-make</id>

  <entry>
    <link rel="alternate" type="application/xml"
          uri="http://starbucks.example.org/order/1234"/>
    <id>http://starbucks.example.org/order/1234</id>
    ...
  </entry>
  ...
</feed>
```

Figure 13 Atom feed for drinks to be made

Starbucks is a busy place and the Atom feed at /orders is updated frequently, so the barista will need to poll it to stay up to date. Polling is normally thought of as offering low scalability; the Web, however, supports an extremely scalable polling mechanism – as we'll see shortly. And with the sheer volume of coffees being manufactured by Starbucks every minute, scaling to meet load is an important issue.

We have two conflicting requirements here. We want baristas to keep up-to-date by polling the order feed often, but we don't want to increase the load on the service or unnecessarily increase network traffic. To avoid crushing our service under load, we'll use a reverse proxy just outside our service to cache and serve frequently accessed resource representations, as shown in Figure 14.

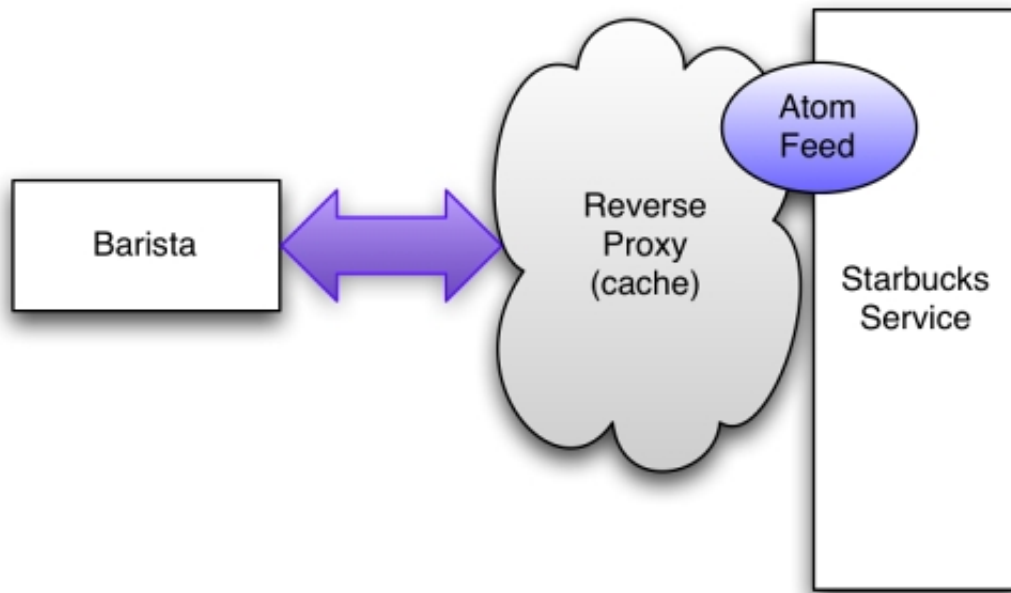


Figure 14 Caching for scalability

For most resources – especially those that are accessed widely, like our Atom feed for drinks – it makes sense to cache them outside of their host services. This reduces server load and improves scalability. Adding Web caches (reverse proxies) to our architecture, together with caching metadata, allows clients to retrieve resources without placing load on the origin server.

A positive side effect of caching is that it masks intermittent failures of the server and helps crash recovery scenarios by improving the availability of resource state. That is, the barista can keep working even if the Starbucks service fails intermittently since the order information will have been cached by a proxy. And if the barista forgets an order (crashes) then recovery is made easier because the orders are highly available.

Of course, caching can keep old orders around longer than needed, which is hardly ideal for a high-throughput retailer like Starbucks. To make sure that cached orders are cleared, the Starbucks service uses the Expires header to declare how long a response can be cached. Any caches between the consumer and service (should) honour that directive and refuse to serve stale orders⁴, instead forwarding the request onto the Starbucks service, which has up-to-date order information.

The response in Figure 13 sets the Expires header on our Atom feed so that drinks turn stale 10 seconds

into the future. Because of this caching behaviour, the server can expect at most 6 requests per minute, with the remainder handled by the cache infrastructure. Even for a relatively poorly performing service, 6 requests per minute is a manageable workload. In the happiest case (from Starbucks' point of view) the barista's polling requests are answered from a local cache, resulting in no increased network activity or server load.

In our example, we use only one cache to help scale-out our master coffee list. Real Web-based scenarios, however, may benefit from several layers of caching. Taking advantage of existing Web caches is critical for scalability in high volume situations.

The Web trades latency for massive scalability. If you have a problem domain that is highly sensitive to latency (e.g. foreign exchange trading), then Web-based solutions are not a great idea. If, however, you can accept latency in the order of seconds, or even minutes or hours, then the Web is likely a suitable platform.

Now that we've addressed scalability, let's return to more functional concerns. When the barista begins to prepare our coffee, the state of the order should change so that no further updates are allowed. From the point of view of a customer, this corresponds to the moment we're no longer allowed to PUT updates of our order (as in Figure 6, Figure 7, Figure 8, Figure 9, and Figure 10).

Fortunately there is a well-defined protocol that we can use for this job: the Atom Publishing Protocol (also known as APP or AtomPub). AtomPub is a Web-centric (URI-based) protocol for managing entries in Atom feeds. Let's take a closer look at the entry representing our coffee in the /orders Atom feed.

```
<entry>
  <published>2008-06-10T19:18:43Z </published>
  <updated>2008-06-10T19:20:32Z</updated>
  <link rel="alternate" type="application/xml"
    uri="http://starbucks.example.org/order/1234"/>
  <id>http://starbucks.example.org/order/1234</id>
  <content type="text+xml">
    <order xmlns="http://starbucks.example.org/">
      <drink>latte</drink>
      <additions>shot</additions>
      <cost>4.00</cost>
    </order>
    <link rel="edit"
      type="application/atom+xml"
      href="http://starbucks.example.org/order/1234/>
    ...
  </content>
</entry>
```

Figure 15 Atom entry for our coffee order

The XML in Figure 15 is interesting for a number of reasons. First, there's the Atom XML, which distinguishes our order from all the other orders in the feed. Then there's the order itself, containing all the information our barista needs to make our coffee – including our all-important extra shot! Inside the order entry, there's a link element that declares the edit URI for the entry. The edit URI links to an order resource that is editable via HTTP. (The address of the editable resource in this case happens to be the same address as the order resource itself, but it need not be.)

When a barista wants to change the state of the resource so that our order can no longer be changed, they interact with it via the edit URI. Specifically they PUT a revised version of the resource state to the edit URI, as shown in Figure 16.

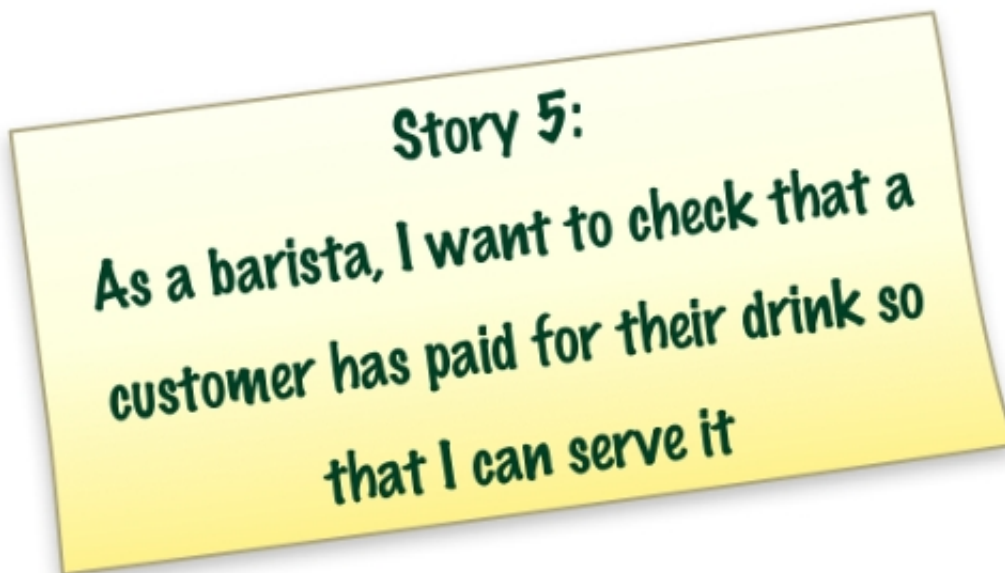
```
PUT /order/1234 HTTP 1.1
Host: starbucks.example.com
Content-Type: application/atom+xml
Content-Length: ...

<entry>
  ...
  <content type="text+xml">
    <order xmlns="http://starbucks.example.org/">
      <drink>latte</drink>
      <additions>shot</additions>
      <cost>4.00</cost>
      <status>preparing</status>
    </order>
    ...
  </content>
</entry>
```

Figure 16 Changing the order status via AtomPub

Once the server has processed the PUT request in Figure 16, it will reject anything other than GET requests to the /orders/1234 resource.

Now that the order is stable the barista can safely get on with making the coffee. Of course, the barista will need to know we've paid for the order before they release the coffee to us, so before handing the coffee over, the barista checks to make sure we've paid. In a real Starbucks, things are a little different: there are conventions, such as paying as you order, and other customers hanging around to make sure you don't run off with their drinks. But in our computerised version it's not much additional work to add this check, and so onto our penultimate story card:



The barista can easily check the payment status by GETting the payment resource using the payment URI in the order.

In this instance the customer and barista know about the payment resource from the link embedded in the order representation. But sometimes it's useful to access resources via URI templates.

URI templates are a description format for well-known URIs. The templates allow consumers to vary parts of a URI to access different resources.

A URI template scheme underpins Amazon's S3 storage service. Stored artefacts are manipulated using the HTTP verbs on URIs created from this template:

`http://s3.amazonaws.com/{bucket_name}/{key_name}`.

It's easy to infer a similar scheme for payments in our model so that baristas (or other authorised Starbucks systems) can readily access each payment without having to navigate all orders: `http://starbucks.example.org/payment/order/{order_id}`

URI templates form a contract with consumers, so service providers must take care to maintain them even as the service evolves. Because of this implicit coupling some Web integrators shy away from URI templates. Our advice is to use them only where inferable URIs are useful and unlikely to change.

An alternative approach in our example would be to expose a `/payments` feed containing (non-inferable) links to each payment resource. The feed would only be available to authorised systems.

Ultimately it is up to the service designer to determine whether URI templates are a safe and useful shortcut through hypermedia. Our advice: use them sparingly!

Of course, not everyone is allowed to look at payments. We'd rather not let the more creative (and less

upstanding) members of the coffee community check each-others' credit card details, so like any sensible Web system, we protect our sensitive resources by requiring authentication.

If an unauthenticated user or system tries to retrieve the details of a particular payment, the server will challenge them to provide credentials, as shown in Figure 17.

Request	Response
GET /payment/order/1234 HTTP 1.1 Host: starbucks.example.org	401 Unauthorized WWW-Authenticate: Digest realm="starbucks.example.org", qop="auth", nonce="ab656...", opaque="b6a9..."

Figure 17 Unauthorised access to a payment resource is challenged

The 401 status (with helpful authentication metadata) tells us we should try the request again, but this time provide appropriate credentials. Retrying with the right credentials (Figure 18), we retrieve the payment and compare it with the resource representing the total value of the order at <http://starbucks.example.org/total/order/1234>.

Request	Response
GET /payment/order/1234 HTTP 1.1 Host: starbucks.example.org Authorization: Digest username="barista joe" realm="starbucks.example.org" nonce="..." uri="payment/order/1234" qop=auth nc=00000001 cnonce="..." reponse="..." opaque="..."	200 OK Content-Type: application/xml Content-Length: ... <payment xmlns="http://starbucks.example.org/"> <cardNo>123456789</cardNo> <expires>07/07</expires> <name>John Citizen</name> <amount>4.00</amount> </payment>

Figure 18 Authorised access to a payment resource

Once the barista has prepared and dispatched the coffee and collected payment, they'll want to remove the completed order from the list of outstanding drinks. As always we'll capture this as a story:



Because each entry in our orders feed identifies an editable resource with its own URI, we can apply the HTTP verbs to each order resource individually. The barista simply DELETES the resource referenced by the relevant entry to remove it from the list, as in Figure 19.

Request	Response
DELETE /order/1234 HTTP 1.1 Host: starbucks.example.org	200 OK

Figure 19 Removing a completed order

With the item DELETED from the feed, a fresh GET of the feed returns a representation without the DELETED resource. Assuming we have well behaved caches and have set the cache expiry metadata sensibly, trying to GET the order entry directly results in a 404 Not Found response.

You might have noticed that the Atom Publishing Protocol meets most of our needs for the Starbucks domain. If we'd exposed the /orders feed directly to customers, customers could have used AtomPub to publish drinks orders to the feed, and even change their orders over time.

Evolution: A fact of Life on the Web

Since our coffee shop is based around self-describing state machines, it's quite straightforward to evolve the workflows to meet changing business needs. For example Starbucks might choose to offer a free Internet promotion shortly after starting to serve coffee:

- July – Our new Starbucks shop goes live offering the standard workflow with the state transitions and representations that we've explored throughout this article. Consumers are interacting with the service with these formats and representations in mind.
- August – Starbucks introduces a new representation for a free wireless promotion. Our coffee workflow will be updated to contain links providing state transitions to the offer. Thanks to the magic of URIs, the links may be to a 3rd party partner just as easily as they could be to an internal Starbucks

resource

```
...  
<next xmlns="http://example.org/state-machine"  
  rel="http://wifi.example.org/free-offer"  
  uri="http://wifi.example.com/free-offer/order/1234"  
  type="application/xml"/>  
...
```

Because the representations still include the original transitions, existing consumers can still reach their goal, though they may not be able to take advantage of the promotion because they have not been explicitly programmed for it.

- September – Consumer applications and services are upgraded so that they can understand and use the free Internet promotion, and are instructed to follow such promotional transitions whenever they occur.

The key to successful evolution is for consumers of the service to anticipate change by default. Instead of binding directly to resources (e.g. via URI templates), at each step the service provides URIs to named resources with which the consumer can interact. Some of these named resources will not be understood and will be ignored; others will provide known state transitions that the consumer wants to make. Either way this scheme allows for graceful evolution of a service while maintaining compatibility with consumers.

The Technology you're about to enjoy is extremely hot

Handing over the coffee brings us to the end of the workflow. We've ordered, changed (or been unable to change) our order, paid and finally received our coffee. On the other side of the counter Starbucks has been equally busy taking payment and managing orders.

We were able to model all necessary interactions here using the Web. The Web allowed us to model some simple unhappy paths (e.g. not being able to change an in process order or one that's already been made) without us having to invent new exceptions or faults: HTTP provided everything we needed right out of the box. And even with the unhappy paths, clients were able to progress towards their goal.

The features HTTP provides might seem innocuous at first. But there is already worldwide agreement and deployment of this protocol, and every conceivable software agent and hardware device understands it to a degree. When we consider the balkanised adoption of other distributed computing technologies (such as WS-) we realise the remarkable success that HTTP has enjoyed, and the potential it releases for system-to-system integration.*

The Web even helped non-functional aspects of the solution. Where we had transient failures, a shared understanding of the idempotent behaviour of verbs like GET, PUT and DELETE allowed safe retries; baked-in caching masked failures and aided crash recovery (through enhanced availability); and HTTPs and HTTP Authentication helped with our rudimentary security needs.

Although our problem domain was somewhat artificial, the techniques we've highlighted are just as applicable in traditional distributed computing scenarios. We won't pretend that the Web is simple (unless

you are a genius), nor do we pretend that that it's a panacea (unless you are an unrelenting optimist or have caught REST religion), but the fact is that the Web is a robust framework for integrating systems at local, enterprise, and Internet scale.

Acknowledgements

The authors would like to thank Andrew Harrison of Cardiff University for the illuminating discussions around “conversation descriptions” on the Web.

About the Authors

Dr. Jim Webber is director of professional services for ThoughtWorks where he works on dependable distributed systems architecture for clients worldwide. Jim was formerly a senior researcher with the UK E-Science programme where he developed strategies for aligning Grid computing with Web Services practices and architectural patterns for dependable Service-Oriented computing and has extensive Web and Web Services architecture and development experience. As an architect with Hewlett-Packard, and later Arjuna Technologies, Jim was the lead developer on the industry's first Web Services Transaction solution. Jim is an active speaker and is invited to speak regularly at conferences across the globe. He is an active author and in addition to "Developing Enterprise Web Services - An Architect's Guide" he is working on a new book on Web-based integration. Jim holds a B.Sc. in Computing Science and Ph.D. in Parallel Computing both from the University of Newcastle upon Tyne. His blog is located at <http://jim.webber.name>.

Savas Parastatidis is a Software Philosopher, thinking about systems and software. He investigates the use of technology in eResearch and is particularly interested in Cloud Computing, knowledge representation and management, and social networking. He's currently with Microsoft Research's External Research team. Savas enjoys blogging at <http://savas.parastatidis.name>.

[Ian Robinson](#) helps clients create sustainable service-oriented capabilities that align business and IT from inception through to operation. He has written guidance for Microsoft on implementing service-oriented systems with Microsoft technologies, and has published articles on consumer-driven service contracts and their role in the software development lifecycle - most recently in [The ThoughtWorks Anthology](#) (Pragmatic Programmers, 2008) and [elsewhere on InfoQ](#). He speaks regularly at conferences on subjects that include RESTful enterprise development and the test-driven foundations of service-oriented delivery.

1 An ETag (an abbreviation of Entity Tag) is a unique identifier for the state of a resource. An ETag for a resource is typically an MD5 checksum or SHA1 hash of that resource's data.

2 We'll see how authentication works from Starbucks' point of view later.

3 When safety is at stake, of course, we just prevent things from going too far wrong in the first place! But receiving coffee isn't a safety critical task, even though it might seem that way for our co-workers most mornings!

4 HTTP 1.1 offers some useful request directives, including max-age, max-stale, and max-fresh, which

allow the client to indicate the level of staleness it is prepared to accept from a cache.

- Sections
- [Enterprise Architecture](#)
- [Architecture & Design](#)
- Topics
- [Patterns](#)
- [Enterprise Architecture](#)
- [Web Services](#)
- [Design Guideline](#)
- [REST](#)
- [Workflow / BPM](#)
- [Business Process Management](#)
- [SOA](#)
- [Design](#)
- [Architecture](#)

Related Editorial

[Does BPM-in-the-Cloud Require RESTful Services? ZapThink Says Yes, but Doubts Exist.](#)

[Is REST the future for SOA?](#)

[The Diary of a REST “Convert”](#)

[WSO2 API Manager 1.4.0 Introduces Multi-Tenancy](#)

Hello stranger!

You need to [Register an InfoQ account](#) or [Login](#) to post comments. But there's so much more behind being registered.

Get the most out of the InfoQ experience.

Tell us what you think

Allowed html: a,b,br,blockquote,i,li,pre,u,ul,p

☐ Email me replies to any of my messages in this thread

[Post Message](#)

Community comments [Watch Thread](#)

Very well done by Ray Krueger Posted 03/10/2008 05:52

"Advanced" REST by Manish Bhatt Posted 03/10/2008 08:14

Excellent work by Oyku Gencay Posted 04/10/2008 04:04

Two Questions by Stephen Molitor Posted 05/10/2008 11:58

Re: Two Questions by Ian Robinson Posted 05/10/2008 05:16

Excellent article gents... by Christian MacLean Posted 06/10/2008 11:13

Multipart vs. Atom vs. your own format by Bill Burke Posted 06/10/2008 11:26

One nit by Mark Baker Posted 07/10/2008 11:28

A few things by Mark Nottingham Posted 14/10/2008 09:36

Re: A few things by Jim Webber Posted 21/10/2008 06:14

A Restful BPM engine in SaaS mode by Matthieu Hug Posted 27/10/2008 05:04

Good but not enough by Alexis Brouard Posted 27/10/2008 11:46

Re: Good but not enough by Tiberiu Fustos Posted 16/11/2008 01:01

Great Article - Few Comments by Colin Jack Posted 03/11/2008 11:05

Re: Great Article - Few Comments by Colin Jack Posted 04/11/2008 05:34

application state vs. resource state by Bob Ferris Posted 28/01/2011 10:21

Very well done 03/10/2008 05:52 by "Ray Krueger"

You even mixed in some OPTIONS action, very nice. Now, if only we had a working sample :)

- [Reply](#)
- [Back to top](#)

"Advanced" REST 03/10/2008 08:14 by "Manish Bhatt"

Very good presentation of some of the advanced aspects of RESTful design.
Good explanation of "ST" of REST.

- [Reply](#)
- [Back to top](#)

Excellent work 04/10/2008 04:04 by "Oyku Gencay"

Great article. A very concise and non holistic approach to RESTful design. Looking forward to more like this. This is a must read.

- [Reply](#)
- [Back to top](#)

Two Questions 05/10/2008 11:58 by "Stephen Molitor"

Great article. It really helped me understand the notion of "hypermedia as the engine of application state".
Two questions:

1. After paying for the order, what would the next state for the customer be and how would he/she get there? In a real Starbucks the barista would notify the customer via a push - "Roy's drink is ready" and the customer would pick it up. But that's not very RESTful. Instead would the customer poll a feed of drinks that are ready? And perhaps the feed could have a URL telling them where to pick up their drink(s).

2. I was a little confused by your use of the 'rel' tag. After placing an order the customer is directed to the payment URL via the 'next' tag. The xml namespace describing the semantics of the next tag points to example.org/state-machine. However the 'rel' tag points to starbucks.example.org/payment for the definition of the payment value. I'm used to having all the possible values of the rel tag defined in the 'next' tag's xmlns, and the rel tag having one of those simple values like "payment", "alternate", etc. Either example.org defines the complete semantics of 'next', or if starbucks has extended it both namespaces should be mentioned in the next tag. At least that's how I understood things but maybe I'm wrong. Anyway this is a minor detail.

- [Reply](#)
- [Back to top](#)

Re: Two Questions 05/10/2008 05:16 by "Ian Robinson"

Hi Stephen

Thanks for your questions. A partial reply to your first only, for now:

I think, as you suggest, some form of client polling will work here. But note that the barista doesn't so much "push" the finished coffee as announce or publish its being completed: it's up to the customer then to fetch the coffee. In this way, some of the reliability issues surrounding the delivery of the coffee and the completion of the transaction are delegated to the customer. If the customer's still in the store and in earshot, he or she can act on the notification and get the coffee: there's no need for some expensive table service (bus) to guarantee delivery from the barista to the customer.

We could have the barista publish to some sort of feed of completed orders, an Atom feed for example, with each customer then polling the feed at intervals in order to determine the status of his or her order. Customers in this particular scenario don't require subsecond updates, but they might reasonably expect to learn their coffee's ready within say 10 seconds of its being completed. As George Malamidis has [pointed out](#), many clients polling a feed every 10 seconds or so poses some pretty significant challenges.

I don't think an Atom feed is appropriate here. A feed of events is useful when the feed consumers are interested in all or most of the events. But here, because they're only interested in the status of their own particular order, each customer will likely have to parse the feed and discard the majority of entries. (But see the FriendFeed update feed and [Simple Update Protocol](#) for an example of using feeds to publish events to clients, each of which is interested in only a fraction of the events.)

Better to create an order status resource per order. Customers can GET (or conditionally GET) a representation of this resource at frequent intervals, and the barista can PUT to it as he or she makes the coffee. We let the customer know about this resource when they first place their order by returning 202 Accepted and setting the Location response header to the URI of the new order status resource.

We can create this order status resource wherever we like. That is, we can shard the order status URI space, with particular shards servicing particular geographies. Better still, we can cache each order status and use something like Mark Nottingham's [cache channels](#) to control the freshness of these responses in a reasonably fine-grained manner. We might specify a max-age of only 5 seconds, but using cache channels, a cache can prolong the freshness of an order status as long as:

- 1) it (the cache) continues polling the cache channel at least as often as the cache channel's precision, and
- 2) the cache doesn't receive a stale event related to that resource from the cache channel.

Note that cache channels use Atom feeds for communication between the origin server and the cache or cache hierarchy - which makes perfect sense since the cache is interested in all or at least a significant portion of events coming off the channel.

The result: many instances of customers hitting the caches, and many fewer instances of caches hitting the origin servers, either to revalidate stale responses or poll cache channels.

- [Reply](#)
- [Back to top](#)

Excellent article gents... 06/10/2008 11:13 by "Christian MacLean"

and very timely!

CM

- [Reply](#)
- [Back to top](#)

Multipart vs. Atom vs. your own format 06/10/2008 11:26 by "Bill Burke"

The values of Atom haven't really clicked with me yet. In this particular example how is it better than something like "multipart/*". On your client or server, you have the added complexity of having to support the Atom XML interaction format. With multipart, we can have the same information(through Location, Content-Location, and Content-Type headers) in a much more condensed format.

Even better than multipart, why not just send back a comma delimited list of order URIs?

One of the things (but not the only thing) that attracted me to REST was that you could focus on the data format you were exchanging between your services and not tunnel your interactions with a intermediary protocol. So far, Atom to me is just a sexier replacement to SOAP.

BTW, I liked the article.

--

Bill Burke

JBoss, a division of Red Hat

- [Reply](#)
- [Back to top](#)

One nit 07/10/2008 11:28 by "Mark Baker"

Nicely done, but application/xml isn't the media type you're looking for.

- [Reply](#)
- [Back to top](#)

A few things 14/10/2008 09:36 by "Mark Nottingham"

Great article, very good illustration. A few points:

* It would be good to explicitly point out that the representations are just that -- representations of state. I've seen many services where the author feels the need to wrap things with "RequestFoo" and "ResponseFoo"

elements, or "Message" elements.

* You say 404 means that the service is too lazy or secure to give a real reason for denying the request. Using 400 and 404 is more appropriate for those, respectively; 404 has defined semantics that some client software will act upon.

* Using OPTIONS seems a bit contrived here, and it should be pointed out that it isn't required; the client is allowed to just try the request and see if it succeeds.

* Expect/Continue is really for cases where you have a long request body and want to see if the URI, method and headers are acceptable before sending the whole thing. It also has some interop problems (i.e., it isn't widely well-supported), so again it's probably easier to just send the request. I understand you're trying to illustrate a point here, but again it's contrived and doesn't get to the heart of the real utility of E/C.

* You update an order with a PUT. My notes on the printout say "No, that's f**ked". This is a flagrant abuse of the semantics of PUT; if you want to combine a representation with the existing state of the resource, use POST or the emerging PATCH.

* You ask the service if it's still possible to PUT using OPTIONS. Again, this leads people to believe that they should OPTIONS before every request, which isn't good practice, or necessary. Furthermore, it's a poor indicator of whether the PUT will be allowed, because if you're updating state that quickly, it very well may change between requests. General consensus is that OPTIONS metadata isn't that fine-grained in time.

* You model your payment resources as /payment/order/1234; why not /order/1234/payment, to leverage relative URIs?

* It would be really nice to see the service interacting with a link that the client gives it (e.g., to a bank account for payment). A lot of service authors make the mistake of only defining the interfaces they provide, rather than the ones that they consume; playing both roles is much more powerful.

* Note that Expires isn't the only mechanism that will keep the order list up-to-date here; POST to the order list will invalidate any intervening caches (as per RFC2616 section 13.10). Unfortunately, this isn't widely implemented, but we've now got it in Squid 2.HEAD, and it should be in release tarballs soon. At Y! we've also been working on communicating invalidations generated in this fashion between peers to keep a cluster in sync, and doing stuff with the grouping mechanism to let them invalidate other URIs. But I digress.

* Intermediaries are one of the more powerful actors in this kind of deployment; not only for caching/acceleration, but also load balancing, routing, and other services. It would be nice to illustrate that (perhaps with an extra barista?).

* Webber, you wrote this whole article just so you could say "The Technology you're about to enjoy is extremely hot", didn't you? :)

- [Reply](#)

- [Back to top](#)

Re: A few things 21/10/2008 06:14 by "Jim Webber"

Dear mnot,

You bet I did :-)

Jim

PS - Partial PUT is all I've got...

- [Reply](#)
- [Back to top](#)

A Restful BPM engine in SaaS mode 27/10/2008 05:04 by "Matthieu Hug"

Hello

Interesting paper; I'd like to share a practical experience on this kind of approach. At RunMyProcess (www.runmyprocess.com) we propose a fully featured BPM engine which technically relies on ReST+Atom (this combination is extremely powerful IMHO). We started to build it 2 years ago, and it's been GA since June 2008. This platform is delivered as a "BPM as a service", meaning it is a "software as a service" platform.

Beyond ideology, what does ReST+Atom bring to a BPM engine? From the user point of view, mainly an extreme ease of integration. A process can be started by a standard HTML form (it's just a post); any data configured within the engine or computed by the engine can be added in any portal, they're all just "RSS" feeds. This last point includes workflow tasks, process monitoring reports, real time BAM data (business activity monitoring), etc. Just feeds; besides, imagine that a feed is a resource like any other, meaning they can be used during process execution: typically useful to build control process in a couple of hours (minutes when you get skilled ;-).

Besides during execution we allow interactions with ReST "services", of course, but also with WS-SOAP, FTP servers or email smtp/pop. From the user point of view all the technical complexity is hidden, they become all equivalent, embedded into a restful process.

ReST+Atom is the technical stack which will contribute to democratize BPM: together with BPMN and "as a service" model, it should put an end to over complicated and endless BPM projects and large spendings.

That combination also opens the door for a largely new spectrum of BPM projects: multientreprise business process.

Enough blah-bah... You're all welcome to register and try, trial is fully featured and free ;-)
(www.runmyprocess.com)

- [Reply](#)
- [Back to top](#)

Good but not enough 27/10/2008 11:46 by "Alexis Brouard"

Hi Jim, Savas and Ian,

As far as I know, it is the first and most complete article on how to use REST and this is cool!
So first, thank you for describing such a sample!

I have however some points to address you.

About the *next* links system, how to handle more complex (and more real-life-business-like) cases like:

- - * When you have a split with multiple paths to follow concurrently after one transition? (note this would have not much sense with the lifecycle of one resource but here we have three resources managed in the same workflow: order, payment and drink)
For instance, we could imagine that payment of the drink would set the order in the *paid* state (that's the process you've described) but also set the myStarbucks account in the *created* state (if the first order at Starbucks would automatically create a myStarBucks account for further loyalty or partnership offers)
- - * When you have a long-running process (which is not the case here as processes you've described all complete in a few minutes), how to ensure new *next* links will not set resources in odd states? (by long-running process, I mean several days, weeks, months or even years and versioning is a tough issue with such processes)
For instance, if we imagine a coffee takes some days to be ready (ok, it's not really realistic in the case of coffee but let's just imagine that as many real-life business processes act this way), if we change the process to enforce customers to fill in a customer service survey after having paid and before being able to receive its drink (i.e. the *next* link appearing on the payment resource points no more to *drink received* state but to a *survey completed* state)

For the last *next* link (the free-offer link, should be figure 20...), from which resource is it extracted? Does this link bounds to the order resource or the payment resource?

If it bounds to the order resource, is there a way to determine programmatically which *next* link to follow? (the payment link or the free offer one)

If it bounds to the payment resource, will the barista be able to see this link when checking a customer payment? (and if the barista was not human, how to avoid him to follow to this - irrelevant - *next* link?)

The latter introduces issues about automation.

In the article, all is clear: we know what to do to go from one state to the other, we understand a conflict root cause, we could even understand a unknown state - a myStarbucks account created or a customer

survey complete -.

But this is barely because we are humans and we natively understand the semantics of the process (as we know the process very well, at least, from a customer point of view).

What if we were programs? (i.e. a Java/JEE or PHP web application - I know, stating a such thing does not really promotes us ;-))

By this I mean, how distributed programs (and a program has a limited "smartness" capability) could conclude and act the same way we can do?

To take another sample than the previous one (aka the barista who should not follow the next link embedded in the resource it gets):

When updating an order, if there is a conflict and if I am a program, I can conclude the following:

- * The difference between what I want to PUT and what exists is a XML element *additions*
- * Since there is a conflict, I cannot PUT my *additions* XML element

And that's all! I cannot conclude on what to do to go further: should I drop my update request (which is the case in the article) or should I send another update (a different addition) because I used a value for the *additions* element that cannot be proceed at this time? (the latter means that the additions we want is invalid only because of the actual state of the resource (and not for all states) and even if it was semantically valid) Is there a way to accurately handle the discovered on the fly *next* link system in order to let a computer use it without blowing our resources out?

In the 3rd story, you indicate in this model, state-machines and workflows are discovered and described on the fly and, because agreements have been made on the semantics (i.e. the set of transitions and states to handle - whether now or in several years -), clients will always know what to do with every state.

I completely agree that semantics must be agreed upfront and, in your sample, processes are discovered - by the client - on the fly.

But I strongly disagree about the fact that processes are described on the fly! Even if they are not described in WS-BPEL or WS-CDL (or in a Java bean or in a C program or whatever), they are however described upfront (at least by your UML diagram).

And the best proof of this is that a set of states has been agreed in the first place (prior any coding). And what is a set of states if not an - underlying - process?

It is then, IMO, sad - as it weakens it - that your article tries to state that no process description is needed upfront.

Moreover, I think that integrating the process in the client (as the client knows all states - not only information it can send or transitions/functions/services it can call - and hold in some way the process description) leads to a big risk: coupling the two.

Coupling is not always a bad thing (there's a lot of coupling in all we've done the past decades and not all has problems). But in the case of processes, especially frequently-changing processes (as the business has to change and adapt itself), coupling processes and clients means that, changing the processes will leads to changing the client (not easy to manage, especially in an automated environment)

Your article seems to get rid of that issue and tend to state that coupling the two is good for every process.

IMO, there's no silver bullet to every problem and it's highly risky to state so (or to not explicitly state the contrary).

Last thing of this long comment: why do the barista set the order status by hand?

In all previous stories in your article, the order doesn't carry its own state as the state-machine does for him! Here, IMO, you left the "Representational State Transfer" (aka REST) principles - as there are no state transfer and no transition used - to go back (in time) to data update (aka the U of CRUD...)

To change the order status to preparing, I expected the barista to PUT a "prepare" resource...

This issue maybe also shows that, in your article, you've mixed cross-resources processes (workflows) and resources lifecycles (which are individual state-machines).

Workflows are not always driven by states (understand workflows are not necessarily state-machines) but lifecycles are always state-machines.

Here, you've described two workflows describing an order-to-drink process and involving three resources (order, payment and drink)

Each of the resource has its own lifecycles (order goes from placed to prepared, payment from tentative to confirmed, drink from ordered to received) which are all distinct from the order-to-drink process (including states when you make your process state-driven).

The order-to-drink process uses resources and makes them evolve (through transitions of their lifecycles) and its state is a combination of resources' states and order-specific information.

I think clearly and cleanly separating all of these is a critical element in REST... And I was disappointed to see it was not in the first and most complete article on how to use REST...

Alexis

- [Reply](#)
- [Back to top](#)

Great Article - Few Comments 03/11/2008 11:05 by "Colin Jack"

Finally we're getting some enterprise oriented REST examples, good stuff. I have a few comments on the content but overall I found it very useful.

The use of PUT on the Order was slightly confusing, you are using it to update a portion of the existing resource and I thought this was a bit of a no-no. Instead I'd have expected a POST (or maybe in the future a PATCH). I realize you have your reasons but I thought it was worth a bit more discussion, especially as an indisciplined client (not sending e-tags or if modified since) could re-POST and get another shot unexpectedly.

The PUT by the barista to update the Order state also seemed fine but I wondered if you'd considered handling this by POSTing a resource (perhaps OrderInProgress) to the Order? Just thinking that going in and updating part of a RESOURCE in this way is quite different to the other examples of RESTful workflow/process solutions that I've seen as they used more of a messaging paradigm (posting messages essentially) which seemed to result in a clearer approach (and allowed us to keep PUT idempotent).

The whole OPTIONS thing also seemed a bit contrived, it's a pity it's so early on in an example like this.

All in all though great stuff.

- [Reply](#)
- [Back to top](#)

Re: Great Article - Few Comments 04/11/2008 05:34 by "Colin Jack"

Just re-reading my comments when I say "re-POST" I obviously mean resend the PUT.

I should also say that I recognise that using PUT to update the Order state is a valid (and maybe preferable) approach, but I'm just not sure how well it works in more complex workflows.

- [Reply](#)
- [Back to top](#)

Re: Good but not enough 16/11/2008 01:01 by "Tiberiu Fustos"

Despite all the heavy-weights commenting on the article so far, I dare to agree with Alexis regarding the mixing of resources (and their states) with processes in the example. The fact that the Order resource contains process information (figure 5) seems suboptimal to me. Ok, the "next step" info is separated in its own namespace (microformat). I think Jean-Jaques' article [The Seven Fallacies of BPM](#) explains the difference pretty well.

Nevertheless, I found the article very instructive and helpful to position REST in the "architecture toolbox" - a silver bullet is it not. Seems to me it still requires a lot of implicit agreements between the parties (see the comments about PUT vs. PATCH vs. POST) - a lot of this goes in the WS-* world in the explicit "contract" in case of cross-enterprise services (in form of WSDLs **and** documents...). Still it makes for a "round" and convincing example. I like [MEST](#) better :-)

- [Reply](#)
- [Back to top](#)

application state vs. resource state 28/01/2011 10:21 by "Bob Ferris"

You somehow mixed the usage of **application state** and resource state in your post. Roy T. Fielding stated once:

"Don't confuse **application state** (the state of the user's application of computing to a given task) with resource state (the state of the world as exposed by a given service). They are not the same thing." (see roy.gbiv.com/untangled/2008/rest-apis-must-be-h...)

That means, GET and HEAD, of course, cause state transitions of the application state. The application e.g., a web browser, will process e.g., render, the retrieved HTML page.

Furthermore, the "additions" have to be add to the "drink" resource, which is part of the order resource and not to the order resource itself.

Another issue is, that the processing model is defined by the specific link type descriptions and not by the referenced resource itself e.g., "anchor elements with an href attribute create a hypertext link that, when selected, invokes a retrieval request (GET) on the URI corresponding to the CDATA-encoded href attribute" (see roy.gbiv.com/untangled/2008/rest-apis-must-be-h...).

The processing models of link types are defined in knowledge representation language specifications. A media type specification is a specific knowledge representation language specification. These specifications themselves can be described with a knowledge representation that is based on a generic knowledge representation structure, i.e. RDF Model - a resource description language. For that matter, every knowledge representation of a resource can be described with knowledge representation languages that are based on such a knowledge representation structure, i.e. Semantic Web ontologies.

To map your flat descriptions on RDF Model based knowledge representations, you can do the following things:

- the resource URI of a description e.g., starbucks.example.com/order/1234, is a subject of an RDF triple
 - the resource type of a description e.g., starbucks.example.com/order, is the type of this subject
 - the value of a "rel" attribute identifies the predicate of an RDF triple e.g., starbucks.example.com/payment; although the example is here a bit inappropriate, because such a relation would rather be named as e.g., "has payment" and hence starbucks.example.com/has_payment, because starbucks.example.com/payment can better identify a resource type (which are currently addressed by a (uninteresting) media type)
 - the value of a "uri" attribute identifies the object of an RDF triple e.g., starbucks.example.com/payment/order/1234; this resource must not exist at the moment of usage (as you also pointed out), we just use it to reference (denote) it
 - the value of a "type" attribute identifies the (intended) range of a property (predicate) or explicitly the resource type of the object resource
- => addressing the available media type (and also resource type) is not really necessary, because a client would request its preferred media types, i.e. a knowledge representation based on the RDF knowledge representation structure can be serialized by applying different serialization formats e.g., Turtle, JSON or XML. However, the necessary semantics are more included in the knowledge representation structure and knowledge representation languages that are based on such a knowledge representation structure, rather than in the serialization syntax, which is, however, also important in the processing chain of machines.

To sum up, I can apply a generic resource description language (knowledge representation structure), like RDF Model, to provide the basis for semantically rich descriptions of media type specifications and further knowledge representation languages. This includes also the description of link types and their processing models.

PS: the order description does not contain all information our barista needs to make our coffee, i.e. if he/she don't know how to make coffee, I should provide a further resolvable reference that leads to a description on how to make coffee, which can be requested, if needed (knowledge discovery on-the-fly) ;)

- [Reply](#)
- [Back to top](#)

[Close](#)

by

on

- View
- [Reply](#)
- [Back to top](#)

[Close](#)

Subject Your Reply

[Quote original message](#)

Allowed html: a,b,br,blockquote,i,li,pre,u,ul,p

☐ Email me replies to any of my messages in this thread

[Close](#)

Subject Your Reply

Allowed html: a,b,br,blockquote,i,li,pre,u,ul,p

☐ Email me replies to any of my messages in this thread

[Close](#)

- Popular
- [Week](#)
- [Month](#)
- [6 months](#)

[How Netflix Deploys Code](#)

[C++11 The Future is Here](#)

[Java Garbage Collection Distilled](#)

[Parallel Keynote: 8 Lines of Code](#)

[Top JavaScript MVC Frameworks](#)

[AngularJS Is Being Ported to Dart](#)

[Interview: Today's Java with Cay Horstmann](#)

Educational Content

- [All](#)
- [Articles](#)
- [Presentations](#)
- [Interviews](#)
- [Books](#)
- [Research](#)



[Jepsen: Testing the Partition Tolerance of PostgreSQL, Redis, MongoDB and Riak](#)

[Kyle Kingsbury](#) Jun 20, 2013



[Building Hypermedia APIs with HTML](#)

[Jon Moore](#) Jun 19, 2013



[Deleting Code at Nokia](#)

[Tom Coupland](#) Jun 19, 2013



[Intro to CLP with core.logic](#)

[Ryan Senior](#) Jun 18, 2013



[Spock: A Highly Logical Way To Test](#)

[Howard Lewis Ship](#) Jun 18, 2013



[Interview and Book Excerpt: Hiring Geeks that Fit](#)

[Shane Hastie](#) Jun 18, 2013

- [Older](#)

Sponsored Links

Infoq Weekly Newsletter

Subscribe to our Weekly email newsletter to follow all new content on InfoQ



**Click to view
how it looks like**

Development

[Xamarin Releases New Guides with Samples Based on Evolve 2013 Training Sessions](#)

[TypeScript Release Adds Generics](#)

[Dart Has Entered Beta with Faster VM, Editor and dart2js](#)

Architecture & Design

[Jepsen: Testing the Partition Tolerance of PostgreSQL, Redis, MongoDB and Riak](#)

[The Uncertain Future of NHibernate](#)

[Deleting Code at Nokia](#)

Process & Practices

[Simplicity for Building the Right Thing](#)

[Martin Fowler at GOTO Amsterdam 2013 about Agile Essence and Fluency](#)

[Spock: A Highly Logical Way To Test](#)

Operations & Infrastructure

[Simplicity for Building the Right Thing](#)

[Jepsen: Testing the Partition Tolerance of PostgreSQL, Redis, MongoDB and Riak](#)

[Yahoo! Open Sources Storm on Hadoop](#)

Enterprise Architecture

[Stratos 2.0 Supports Any Runtime and 30 IaaS](#)

[Linda Rising Talks About Incentives at GOTO Amsterdam 2013](#)

[Yahoo! Open Sources Storm on Hadoop](#)

- [Home](#)
- [All topics](#)
- [QCon Conferences](#)
- [About us](#)
- [About You](#)
- [Our Contributor Team](#)
- [Purpose Index](#)

- [Create account](#)
- [Login](#)
- **EVENTS**
- [QCon New York, June 12-14](#)
- [QCon São Paulo, Aug 30-31](#)
- [QCon Shanghai, Oct 23-25](#)
- [QCon San Francisco, Nov 11-13](#)

Infoq Weekly Newsletter

Subscribe to our Weekly email newsletter to follow all new content on InfoQ

Click to view
an example



- [Your personalized RSS](#)
- [For daily content and announcements](#)
- [For major community updates](#)

Personalize Your Main Interests

- ☒ Development
- ☒ Architecture & Design
- ☒ Process & Practices
- ☒ Operations & Infrastructure
- ☒ Enterprise Architecture

General Feedback Bugs Advertising Editorial
feedback@infoq.com bugs@infoq.com sales@infoq.com editors@infoq.com

InfoQ.com and all content
copyright © 2006-2013
C4Media Inc. InfoQ.com
hosted at [Contegix](#), the best ISP
we've ever worked with.
[Privacy policy](#)