microservices using NGINX Plus, as an ebook – Microservices: From Design to Deployment. Also, please look at the new Microservices Solutions page.

The first article in this seven-part series about designing, building, and deploying microservices introduced the Microservices Architecture pattern. It discussed the benefits and drawbacks of using microservices and how, despite the complexity of microservices, they are usually the ideal choice for complex applications. This is the second article in the series and will discuss building microservices using an API Gateway.

When you choose to build your application as a set of microservices, you need to decide how your application's clients will interact with the microservices. With a monolithic application there is just one set of (typically replicated, load-balanced) endpoints. In a microservices architecture, however, each microservice exposes a set of what are typically fine-grained endpoints. In this article, we examine how this impacts client-to-application communication and proposes an approach that uses an API Gateway.

Introduction

Let's imagine that you are developing a native mobile client for a shopping application. It's likely that you need to implement a product details page, which displays information about any given product.

For example, the following diagram shows what you will see when scrolling through the product details in Amazon's Android mobile application.

- Various recommendations, including other products this product is frequently bought with, other products bought by customers who bought this product, and other products viewed by customers who bought this product
- Alternative purchasing options

When using a monolithic application architecture, a mobile client would retrieve this data by making a single REST call (`GET api.company.com/productdetails/`*`productId`*) to the application. A load balancer routes the request to one of N identical application instances. The application would then query various database tables and return the response to the client.

In contrast, when using the microservices architecture the data displayed on the product details page is owned by multiple microservices. Here are some of the potential microservices that own data displayed on the example product details page:

- Shopping Cart Service – Number of items in the shopping cart
- Order Service – Order history
- Catalog Service – Basic product information, such as its name, image, and price
- Review Service – Customer reviews
- Inventory Service – Low inventory warning
- Shipping Service – Shipping options, deadlines, and costs drawn separately from the shipping provider's API
- Recommendation Service(s) – Suggested items

We need to decide how the mobile client accesses these services. Let's look at the options.

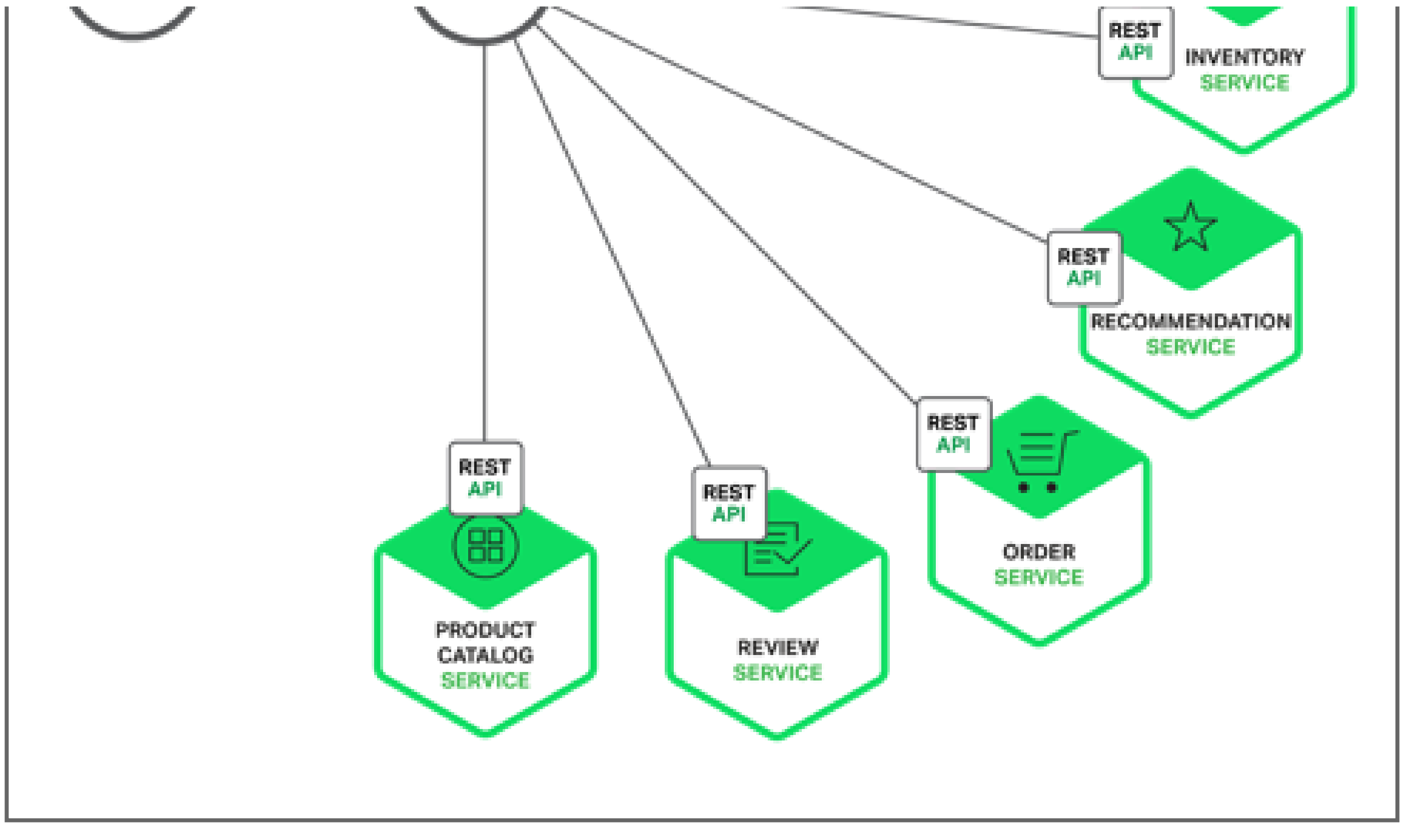
Direct Client-to-Microservice Communication

In theory, a client could make requests to each of the microservices directly. Each microservice would have a public endpoint (**https://*serviceName*.api.company.name**). This URL would map to the microservice's load balancer, which distributes requests across the available instances. To retrieve the product details, the mobile client would make requests to each of the services listed above.

Unfortunately, there are challenges and limitations with this option. One problem is the mismatch between the needs of the client and the fine-grained APIs exposed by each of the microservices. The client in this example has to make seven separate requests. In more complex applications it might have to make many more. For example, Amazon describes how hundreds of services are involved in rendering their product page. While a client could make that many requests over a LAN, it would probably be too inefficient over the public Internet and would definitely be impractical over a mobile network. This approach also makes the client code much more complex.

Another problem with the client directly calling the microservices is that some might use protocols that are not web-friendly. One service might use Thrift binary RPC while another service might use the AMQP messaging protocol. Neither protocol is particularly browser- or firewall-friendly and is best used internally. An application should use protocols such as HTTP and WebSocket outside of the firewall.

Another drawback with this approach is that it makes it difficult to refactor the microservices. Over time we might want to change how the system is partitioned into services. For example, we might merge two services or split a service into two or more services. If, however, clients communicate directly with the services, then performing this kind of refactoring can be extremely difficult.



The API Gateway is responsible for request routing, composition, and protocol translation. All requests from clients first go through the API Gateway. It then routes requests to the appropriate microservice. The API Gateway will often handle a request by invoking multiple microservices and aggregating the results. It can translate between web protocols such as HTTP and WebSocket and web-unfriendly protocols that are used internally.

The API Gateway can also provide each client with a custom API. It typically exposes a coarse-grained API for mobile clients. Consider, for example, the product details scenario. The API Gateway can provide an endpoint (**/productdetails?productid=*xxx***) that enables a mobile client to retrieve all of the product details with a single request. The API Gateway handles the request by invoking the various services – product info, recommendations, reviews, etc. – and combining the results.

an API Gateway.

Implementing an API Gateway

Now that we have looked at the motivations and the trade-offs for using an API Gateway, let's look at various design issues you need to consider.

Performance and Scalability

Only a handful of companies operate at the scale of Netflix and need to handle billions of requests per day. However, for most applications the performance and scalability of the API Gateway is usually very important. It makes sense, therefore, to build the API Gateway on a platform that supports asynchronous, nonblocking I/O. There are a variety of different technologies that can be used to implement a scalable API Gateway. On the JVM you can use one of the NIO-based frameworks such Netty, Vertx, Spring Reactor, or JBoss Undertow. One popular non-JVM option is Node.js, which is a platform built on Chrome's JavaScript engine. Another option is to use NGINX Plus. NGINX Plus offers a mature, scalable, high-performance web server and reverse proxy that is easily deployed, configured, and programmed. NGINX Plus can manage authentication, access control, load balancing requests, caching responses, and provides application-aware health checks and monitoring.

Using a Reactive Programming Model

The API Gateway handles some requests by simply routing them to the appropriate backend service. It handles other requests by invoking multiple backend services and aggregating the results. With some requests, such as a product details request, the requests to backend services are independent of one another. In order to minimize response time, the API Gateway should perform independent requests concurrently. Sometimes, however, there are dependencies between requests. The API Gateway might first need to validate the request by calling an authentication service, before routing the request to a backend service. Similarly, to fetch information about the products in a

Service Discovery

The API Gateway needs to know the location (IP address and port) of each microservice with which it communicates. In a traditional application, you could probably hardwire the locations, but in a modern, cloud-based microservices application this is a nontrivial problem. Infrastructure services, such as a message broker, will usually have a static location, which can be specified via OS environment variables. However, determining the location of an application service is not so easy. Application services have dynamically assigned locations. Also, the set of instances of a service changes dynamically because of autoscaling and upgrades. Consequently, the API Gateway, like any other service client in the system, needs to use the system's service discovery mechanism: either Server-Side Discovery or Client-Side Discovery. A later article will describe service discovery in more detail. For now, it is worthwhile to note that if the system uses Client-Side Discovery then the API Gateway must be able to query the Service Registry, which is a database of all microservice instances and their locations.

Handling Partial Failures

Another issue you have to address when implementing an API Gateway is the problem of partial failure. This issue arises in all distributed systems whenever one service calls another service that is either responding slowly or is unavailable. The API Gateway should never block indefinitely waiting for a downstream service. However, how it handles the failure depends on the specific scenario and which service is failing. For example, if the recommendation service is unresponsive in the product details scenario, the API Gateway should return the rest of the product details to the client since they are still useful to the user. The recommendations could either be empty or replaced by, for example, a hardwired top ten list. If, however, the product information service is unresponsive then API Gateway should return an error to the client.

The API Gateway could also return cached data if that was available. For example, since product prices change infrequently, the API Gateway could return cached pricing data if the pricing service is unavailable. The data can be cached by the API Gateway itself or be stored in an external cache such as Redis or Memcached. By returning either default data or cached data, the API Gateway ensures that system failures do not impact the user experience.

4. Service Discovery in a Microservices Architecture

5. Event-Driven Data Management for Microservices

6. Choosing a Microservices Deployment Strategy

7. Refactoring a Monolith into Microservices

You can also download the complete set of articles, plus information about implementing microservices using NGINX Plus, as an ebook – Microservices: From Design to Deployment.

*For a detailed look at additional use cases, see our three-part blog series, **Deploying NGINX Plus as an API Gateway**:*

- *Part 1 provides detailed configuration instructions for several use cases.*
- *Part 2 extends those use cases and looks at a range of safeguards that can be applied to protect and secure backend API services in production.*
- *Part 3 explains how to deploy NGINX Plus as an API gateway for gRPC services.*

Guest blogger Chris Richardson is the founder of the original CloudFoundry.com, an early Java PaaS (Platform as a Service) for Amazon EC2. He now consults with organizations to improve how they develop and deploy applications. He also blogs regularly about microservices at http://microservices.io.

34 Comments NGINX 1 Login

♡ Recommend 20 Tweet Share Sort by Best

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Paulo Merson • 4 years ago

Excellent article! I'm also a frequent reader of your microservices pattern catalog.

Where you mention an API Gateway endpoint to retrieve product details, you say "The API Gateway handles the request by invoking the various services – … – and combining the results." Invoking services and combining results is what a composition controller service would do. This service contains task specific logic and knows what other services to invoke, how to parse application specific JSON documents to create another application specific JSON, how to handle exceptions, how to demarcate transactions and use compensation, how to dehydrate if one of the services will take long, how to set correlation identifiers and handle callbacks if needed. Does an API Gateway do any of that?!

I thought an API Gateway provided interceptor-like functionality (routing, transformation, protocol bridging, analytics, security controls). Invoking services and combining results is more like what you would do in an orchestrator service. Is the API Gateway already incorporating orchestration server features?

5 ^ | ⌄ • Reply • Share ›

Eric Duncan ➦ Paulo Merson • 3 years ago

In a typical design, you are only making 2 or 3 calls from the API Gateway to the backend components:

coming back in potentially to a new api server.

This is problematic as a whole as it creates IO bottlenecks.

The solution is to use API abstraction and shared IO through the new API pattern which is designed for distributed architectures unlike the OLD API pattern which was designed in the 70's and not intended for distributed architectures.

⌃ | ⌄ • Reply • Share ›

Siri → Eric Duncan • 3 years ago

Hi, we have been stuck with the problem of choosing good software that provide orchestrator services. Wondering if you had any recommendations.

⌃ | ⌄ • Reply • Share ›

Eric Duncan → Siri • 3 years ago

There are several, and can depend on your tech stack. Feel free to email as that can be a big convo.

⌃ | ⌄ • Reply • Share ›

Siri → Eric Duncan • 3 years ago

Cool, thanks! Sent you an email just now.

⌃ | ⌄ • Reply • Share ›

kerfufl • 4 years ago

" If you are using the JVM you should definitely consider using Hystrix. And, if you are running in a non-JVM environment, you should use an equivalent library." So, I'd love to become aware of some non-JVM equivalents for this circuit-breaker pattern. Anyone implemented this using Nginx?

4 ⌃ | ⌄ • Reply • Share ›

Aldric • 3 years ago • edited

For those who want a example of what is implementing micro services in environment like nodeJs you can watch the video

The API Gateway is like an ESB is an interesting argument.
On the one hand, it is the central point of contention.
Potentially different teams have to contribute code and/or configuration metadata to the gateway to expose their functionality.
On the other hand, the API gateway might be able to dynamically discovery much of the routing information, which alleviates that problem.
Also, the API Gateway is on the edge of the system - its model(s) are at the API layer rather than being an integral part of the application's business logic.

Something for you to consider: if you don't use an API gateway then how do clients of the system communicate with the system?

1 ^ | ⌄ • Reply • Share ›

Robert Karczewski ➦ Chris Richardson • 4 years ago

Maybe it is all about naming but, I think that API Gateway in the context is back-end for front-end couse as I understand there should be dedicated API Gateway for each client. My idea is to give access to the system in three ways:
- UI composition for web application,
- back-end for front-end for mobile app,
- API Exposure Layer (APIGee in this case) for API exposure for 3rd parties.
As the API is going to be exposed as commercial offer of my company the role of API Exposure Layer is more about AAA and protocol conversion than integration of data and services.
I persuaded my business stakeholders to implement MSA with promise of shortening Time To Market not promising high availability, ease of scaling and other MSA benefits because for them it is all IT concerns.
My current environment is mature SOA environment with number of ESBs that slows down delivery of business new ideas.
I'd like to know how to deal with danger of polluting API Gateway with logic of services. Are any patterns how keep functional decoupling in real life implementation of MSA?

Selim Öber ➦ MGG • 4 years ago • edited

You could use something similar to JWT claims (https://jwt.io/) and avoid the backend trip for a configurable period of time (token expiration).

1 ⌃ | ⌄ • Reply • Share ›

anas tayaa • 2 months ago

Thanks for the famous article. My question is, Is there any way to communicate between microservices using Kong API management (without using feign client)

⌃ | ⌄ • Reply • Share ›

CobNinja • 7 months ago

Are there any clear performance benefits of NGiNX plus over node.js? It's hard to find concrete numbers.

⌃ | ⌄ • Reply • Share ›

Faisal Memon ➦ CobNinja • 7 months ago

You can see NGINX Performance numbers here: https://www.nginx.com/resou...

⌃ | ⌄ • Reply • Share ›

hadafnet • 9 months ago

such a great content

please provide and share this tutorial for node js microservices implementation

⌃ | ⌄ • Reply • Share ›

perrohunter • 3 years ago

Do you consider that the API Gateware should server the Web page appliation as well? Or should a separate service serve the web application and the API Gateway live into a subdomain for subsequent calling?

For example: I serve my angular application from example.com (web ui microservice location) and have the angular app call the API Gateway at api.example.com (api gateway microservice location)?

⌃ | ⌄ • Reply • Share ›

service.

If you are thinking A and B are both backend services needing to talk to each other, that's a difference approach. In a microservice setup, you would use inter process communication between services.

The API Gateway just routes more-or-less public API requests to publicly exposed API endpoints on your services. But that's just one opinion. Several setups use those endpoints to talk to others. I find that cross-cutting concerns though as now you are supporting public API endpoints as well as backend private xalla., so I usually keep them separate.

1 • Reply • Share ›

Danny • 4 years ago

Great article. Really helping me to understand API gateway for beginner like me! Can you explain a little bit more on the handling for partial failures if the backend APIs are related to action such as CREATE, UPDATE, DELETE. In some cases, we may rollback those previous executed APIs when exception encountered. But sometimes rollback seems not doable. Any advice?

• Reply • Share ›

Chris Richardson ➦ Danny • 4 years ago

Thanks. Glad you like the article.
I think that generally speaking an Create/Update/Delete request should be handled by a single microservice.
That service should update it's data and then publish events to trigger the update of data owned by other services.
See https://www.nginx.com/blog/...

• Reply • Share ›

Jeff Walker • 4 years ago

We have a situation where we'd like to run some specific Java code for certain requests, possibly before the request reaches a Microservice instance. Is it possible for the NGINX API Gateway to act as an application host, like Tomcat can host web apps. Or is the concept more like, NGINX will route the request to a Tomcat where the request can be pre-processed?

Products

NGINX Plus
NGINX Controller
NGINX Unit
NGINX Amplify
NGINX Web Application Firewall

NGINX on Github

NGINX Open Source
NGINX Unit
NGINX Amplify
NGINX Kubernetes Ingress Controller
NGINX Microservices Reference Architecture
NGINX Crossplane

Connect With Us

STAY IN THE LOOP

Solutions

ADC / Load Balancing
Microservices
Cloud
Security
Web & Mobile Performance
API Management

Resources

Documentation
Ebooks
Webinars
Datasheets
Success Stories
Blog
FAQ
Learn
Glossary

Support

Professional Services
Training
Customer Portal Login

Partners

Amazon Web Services
Google Cloud Platform
IBM
Microsoft Azure
Red Hat
Find a Partner
Certified Module Program

Company

About NGINX
Careers
Leadership
Press
Events