ANALYSIS OF MONOLITHIC AND MICROSERVICE ARCHITECTURES
FEATURES AND METRICS

In this paper the information technologies stack is presented. These technologies are used during network architecture deployment. The analysis of technological advantages and drawbacks under investigation for monolithic and network architectures will be useful during of cyber security analysis in telecom networks. The analysis of the main numeric characteristics was carried out with the aid of Kubectl. The results of a series of numerical experiments on the evaluation of the response speed to requests and the fault tolerance are presented. The characteristics of the of monolithic and microservice-based architectures scalability are under investigation. For the time series sets, which characterize the network server load, the value of the Hurst exponent was calculated.

The research main goal is the monolithic and microservice architecture main characteristics analysis, time series data from the network server accruing, and their statistical analysis.

The methodology of Kubernetes clusters deploying using Minikube, Kubectl, Docker has been used. Application deploy on AWS ECS virtual machine with monolithic architecture and on the Kubernetes cluster (AWS EKS) were conducted.

The investigation results give us the confirmation, that the microservices architecture would be more fault tolerance and flexible in comparison with the monolithic architecture. Time series fractal analysis on the server equipment load showed the presence of long-term dependency, so that we can treat the traffic implementation as a self-similar process.

The scientific novelty of the article lies in the application of fractal analysis to real time series: use of the kernel in user space, kernel latency, RAM usage, caching of RAM collected over 6 months with a step of 10 seconds, establishing a long-term dependence of time series data.

The practical significance of the research is methodology creation of the monolithic and microservice architectures deployment and exploitation, as well as the use of time series fractal analysis for the network equipment load exploration.

Key words: monolith architecture, microservice architecture, replay delay, scalability, fault tolerance, monitoring, time series, Hirst’s exponent.

TATYANA SELIVORSTOVA, SERGEY KLISSH, SERHII KYRYCHENKO, ANTON GUDA, KATERYNA OSTROVSKYAYA
National Metallurgical Academy of Ukraine

ANALIZ OSOBLYWOSTEI TA METRIK MONOLITNOЇ I MIKROSERVISNOЇ ARKHITEKTURY

В роботі представленний стек інформаційних технологій, які засвячуються при розгортанні мережевої архітектури. Представлений аналіз технологічних переваг та недоліків монолітної та мережевої архітектури може бути корисним при розглядах питань безперервної та в телекомунікаційних мережах. Проведений аналіз основних кількісних характеристик засобами Kubectl. Наведені результати серії обчислювальних експериментів по оцінці швидкості відповіді на запити, відмовості, розглянути питання масштабування монолітної та мікросервісної мережевої архітектури. Для наборів часових рядів, що характеризують завантаження мережевого серверу, обчислені значення показника Харста.

Метою роботи є проведення дослідження основних характеристик монолітної та мікросервісної архітектури, збір даних часових рядів з сервера та її статистичний аналіз. Використано методику розгортання Kubernetes кластеру засобами Minikube, Kubectl, Docker. Проведено розгортання додатку на віртуальній машині AWS EC2 – монолітна архітектура та Kubernetes кластеру на AWS EKS – мікросервісна архітектура.

Результати отримані в роботі підтвердили, що архітектура мікросервісу має набагато більшу відмовостійкість та гнучкість у порівнянні з монолітною архітектурою. Фрактальний аналіз часових рядів навантаження серверного обладнання показав наявність довгострокової залежності, що дає змогу представити реалізацію графіку як самоподібний процес.

Наукова новизна роботи полягає у застосуванні фрактального аналізу до реальних часових рядів: використання ядра в просторі користувача, часу очікування ядра, використання оперативної пам’яті, використання оперативної пам’яті, зібрані на протязі 6 місяця з кроком 10 секунд, встановлені довгострокові залежності часинних рядів.

Практична значимість роботи полягає у викладених методології розгортання та експлуатації монолітної та мікросервісної архітектур, застосуванні фрактального аналізу часових рядів для аналізу мережевого навантаження.

Ключові слова: архітектура, монолітна, мікросервісна, затримка, надійність, масштабування, місткість, спритність, моніторинг, часовий ряд, показник Харста.

Introduction

The rapid growth in the number of complex and extensive web-applications has led to the introduction of the latest communication technologies within the project [1, 2]. The monolithic architecture [3, 4] is logically replaced by microservices [5, 6]. In this regard, the task of determining the technical [7, 8] and technological [9, 10] features of this technology [11, 12] becomes relevant.

Related works

Conceptual differences between monolithic and microservice implementations
It is known that the monolithic architecture involves the process of developing the application as a whole. Any changes, even the smallest ones, require significant restructuring and deployment of the entire application [13, 14]. Over time, it becomes increasingly difficult to maintain a clear and understandable modular structure, as changes in the logic of one module tend to affect the code of other modules [15]. Unlike a monolithic architecture, a microservice architecture is an approach to building a server application [16] as a set of almost unrelated services (Figure 1). Services are developed and deployed separately and independently of each other [17]. A separate process is required to run each service. Communication between processes is implemented using HTTP / HTTPS, WebSockets or some other protocols (Figure 2).

The benefits of microservice architecture [19, 20] become apparent when the structure of monolith [21, 22] becomes so complex that it affects both development and maintenance, as well as runtime performance [23, 24]. Microservice architecture does not benefit smaller applications as the overheads due to the distributed architecture both in development and operation outweigh the benefits [25, 26].

Research of the main characteristics of application implementation on microservices and monolith

Creation of test monolithic and microservice architecture

For comparison, we deployed the Wordpress application in an AWS EKS cluster (microservice architecture) and on the virtual machine AWS EC2 (monolithic architecture). We used the computing power of the cloud provider Amazon Web Services and the same type of virtual machines for AWS EKS and AWS EC2. The AWS EKS nodes were created on the basis of m5.large virtual machines. These machines have 2 CPUs and 8 GB of RAM. We also used an Application Load Balancer. It is important to note, that pods with MariaDB and Wordpress itself were deployed on different pods.

Speed of response to a request

Let's check the speed of response to requests for microservice and monolithic architecture [30, 27]. To check, we will use the application "httping", which can determine the speed of responses to http requests. The command "httping -i 1 -c 50" means that we will send 50 http requests with an interval of 1 second.

Run the command for the microservice application:

```bash
httping -i 1 -c 50 http://a179086a04dc94aa7b0fc6a6e80987e-923853456.us-east-1.elb.amazonaws.com/wp-admin/
```

The average response time to a request in a microservice implementation is 24 ms. And at a consecutive call of commands the speed of the answer will depend on length of turn.

Let's now run the same command for the monolithic application.

```bash
httping -i 1 -c 50 http://54.93.175.7/
```

The average response time to a request in a monolith implementation is 1.12 ms.

Thus, the response time to the request is determined by the formula:

\[
L(n) = n \times F ,
\]

where \( n \) is the length of the queue, \( F \) is the delay in processing one request. For microverses \( F = 24 \text{ ms} \), for monolith \( F = 1.12 \text{ ms} \).
Thus, we see that the call delay in the microservice architecture is on average 24 ms, which is almost twenty times the delay of the monolithic architecture. It should be noted that the delay increases significantly as the queue increases.

The monolith does not have a network delay because all calls are local. Microservices cannot stay ahead of physics when it comes to the speed of response to requests.

Reliability of network requests

The following calculations are used to assess the reliability of network requests. Suppose a service accesses another service over a microservice cluster network with 99.9% reliability. This means that out of 1000 calls, one will fail due to network problems.

\[ R(n) = P^n, \]  

where \( n \) is the length of the call chains, \( P \) is the reliability of the call.

Now, if this service calls another service, we get 99.8% fault tolerance. For a chain of calls with a depth of up to 10 calls, we achieve 99% reliability – which means that 1 in 100 calls fails (Figure 5). Because all calls in the monolith are local, there is no chance of a network failure.

Because network problems are an expected phenomenon, microservices offer some solutions. For example, the open source Spring Cloud application offers transparent load balancing and troubleshooting for Java.

In a microservice cluster, services can fail. In this case, the cluster manager simply launches a similar service with the same characteristics. This makes the microservice architecture extremely stable.

Scaling

There are ways to scale the monolith. You can run multiple instances and configure routing. Or you can run multiple threads. For microservice architecture, this is also true. But you can scale microservices with less resources. This means that for the money spent on resources, microservices provide more bandwidth. More accurate scaling is also possible. If the monolith uses all the resources, the way to handle more connections is to output a second instance. If one microservice uses all the resources, only that service will need more copies. Because microservices are less
resource-intensive, it saves resources. Because scaling is simple and accurate, it means that only the required amount of resources is used.

For example, suppose a monolith runs on an m5.16xlarge instance of Amazon Web Service. The m5.16xlarge instance offers a whopping 64 processors and 262 GB of RAM. At the moment, the cost is also enormous – 2211.84 US dollars per month for a round-the-clock virtual machine. For the same money you can buy 9 instances of c5.2xlarge with 8 virtual processors and 16 GB of RAM, each of which works around the clock and without days off. But most workloads do not require full use of resources around the clock. Instead, the use of resources reaches a maximum at certain hours, and at other times the load becomes less.

\[
Pr(H) = A \cdot H_{day} \cdot D, \tag{3}
\]

where \( D \) is the number of days, \( H_{day} \) is the number of operating hours per day, \( A = 3 \) $ is the cost of 1 hour of the 9 instances.

Below is Figure 6, which shows how much these 9 smaller copies cost if they run only for a certain period of time instead of 24/7 (Figure 5).

Because the allocated resources are cheaper than one, which is charged hourly, there is a time when a single copy becomes cheaper. In this example, the transition is (Figure 6):
- 744 hours, if the instances work 100% of the time;
- 1104 hours, if the instances work 67% of the time;
- 1464 hours, if the instances work 50% of the time;
- 2184 hours, if the instances work 34% of the time;
- 4344 hours, if the instances work 17% of the time.

So, over time, using microservices can become more expensive than a monolith. However, microservices allow us to use resources more efficiently and scale more accurately.

The results of comparing the characteristics of the implementation of the application deployment on microservices and monolith

We researched the configuration of the Kubernetes orchestration platform for deploying an application with a microservice architecture. The result of our research is a list of some advantages and disadvantages of both microservice and monolithic architecture (Table 1).

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Microservice architecture</th>
<th>Monolithic architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability</td>
<td>More fault tolerant. Microservices can be easily distributed across multiple nodes. If one node fails, the others can take the load. Microservices are based on the assumption that the network is unreliable.</td>
<td>Simpler, but less fault tolerant. Monoliths have fewer moving parts. Monolithic architecture is suitable for undistributed loads.</td>
</tr>
<tr>
<td>Scaling</td>
<td>Easier and more accurate scaling. Only the components required for load handling are scaled.</td>
<td>Scaling is more complex, slower, and less accurate because almost all components are scaled.</td>
</tr>
<tr>
<td>Capacity</td>
<td>Responds to requests more slowly because it has network nodes and an additional load of images and orchestration.</td>
<td>Responds faster to requests because the monolithic architecture does not have networked nodes, images, and orchestration.</td>
</tr>
<tr>
<td>Agility</td>
<td>More agile. Code updates can be deployed and installed very quickly because microservices are loosely interconnected.</td>
<td>Less agile. Updating the code takes a lot of time and checking before deployment.</td>
</tr>
</tbody>
</table>
Monitoring of network infrastructure

The information technology stack for monitoring network equipment includes Grafana, Prometheus, Loki. Data from services is collected in Prometheus Server and visualized using Grafana. Using Loki allows you to organize and get online access to logs.

The architecture of the project is structured in such a way that a processor load of more than 75 percent is considered unacceptable if the project is working reliably. The metric "Memory load,%" has a threshold value of 80, an exaggeration of which can become critical for the system's performance.

For reliable system operation:
1) planning of cluster large capacities at the stage of project deployment;
2) limiting the number of clients;
3) adaptive increase in cluster capacities.

The first option - excellent, however, rather expensive - involves large material investments at the start of the project.

Limiting the number of clients usually goes against a business strategy.

The adaptive increase in processor power requires flexible management of the available processor resources, which is possible manually (the human factor is a big obstacle to this approach) or by means of Prometheus / Grafana.

On the Grafana side, the "Alert" mechanism is implemented, which launches deployment-patch.yaml, which automatically increases the processor capacity by connecting an additional Node to the micro-environment.

Real time series research of server state

Consider a set of time series of server state in real time. Data were collected over a six-month period in 10 second increments. Table 2 shows the calculations of the Hurst exponent for the time series of server equipment when the window size is varied [31].

The R / S-analysis of changes in the Hurst exponent, presented in Table 2, showed the following characteristic features of these implementations: long-term dependence. The Hurst parameter is significantly greater than 0.5 in all cases, corresponds to a process which is trending (persistent).

Table 2

<table>
<thead>
<tr>
<th>File name</th>
<th>Window size</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>cpu-user.xml</td>
<td>0.79</td>
<td>0.90</td>
</tr>
<tr>
<td>cpu-wait.xml</td>
<td>0.75</td>
<td>0.85</td>
</tr>
<tr>
<td>memory-used.xml</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>memory-cached.xml</td>
<td>0.97</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Conclusion

We investigated the configuration of the microservice and monolithic application architecture. We concluded that the microservice architecture has much greater fault tolerance and flexibility compared to the monolithic architecture. Applications deployed in a microservice architecture are much easier to scale and update than applications deployed in a monolithic architecture. Application developers with a microservice architecture communicate much more efficiently in smaller teams compared to the large teams required to develop applications with a monolithic architecture.

However, the monolithic application deployment architecture is simpler and great for deploying an application with unallocated workloads. Monolithic architecture responds much faster to requests and has fewer moving parts. Therefore, it may be easier to administer and maintain applications with monolithic architecture.

Time series of server equipment load have been investigated by the methods of fractal analysis. The presence of a long-term dependence is shown, which makes it possible to represent the implementation of traffic as self-similar process.

References


<table>
<thead>
<tr>
<th>Name</th>
<th>Position and Department</th>
<th>Email</th>
<th>ORCID / Scopus ID</th>
<th>Additional Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sergey Klishch</td>
<td>postgraduate, Department of information technology and systems, National Metallurgical Academy of Ukraine.</td>
<td><a href="mailto:sergey.klishch@gmail.com">sergey.klishch@gmail.com</a></td>
<td>orcid.org/0000-0002-7799-1004</td>
<td>аспірант кафедри інформаційних технологій та систем, Національна металургійна академія України.</td>
</tr>
<tr>
<td>Serhii Kyrychenko</td>
<td>master, Department of information technology and systems, National Metallurgical Academy of Ukraine.</td>
<td><a href="mailto:skir79@gmail.com">skir79@gmail.com</a></td>
<td>orcid.org/0000-0001-9700-4136</td>
<td>магістр кафедри інформаційних технологій та систем, Національна металургійна академія України.</td>
</tr>
<tr>
<td>Anton Guda</td>
<td>doctor of engineering, associate professor, Department of information technology and systems, National Metallurgical Academy of Ukraine.</td>
<td><a href="mailto:atu.guda@gmail.com">atu.guda@gmail.com</a></td>
<td>orcid.org/0000-0003-1139-1580</td>
<td>доктор технічних наук, доцент, професор кафедри інформаційних технологій та систем, Національна металургійна академія України.</td>
</tr>
<tr>
<td>Kateryna Ostrovskaya</td>
<td>candidate of technical science, assistant professor, Department of information technology and systems, National Metallurgical Academy of Ukraine.</td>
<td><a href="mailto:kuostrovskaya@gmail.com">kuostrovskaya@gmail.com</a></td>
<td>orcid.org/0000-0002-9375-4121</td>
<td>кандидат технічних наук, доцент, доцент кафедри інформаційних технологій та систем, Національна металургійна академія України.</td>
</tr>
</tbody>
</table>