UCLOUD

大型分布式系统设计的一些黄金原则和实例

俞圆圆 a.k.a Y3
YuanYuan.Yu@UCloud.cn

UCLOUD

Agenda

- 4个在生产环境中的1, 2级事故
- 故障排查的过程和根因分析
- 反思和教训

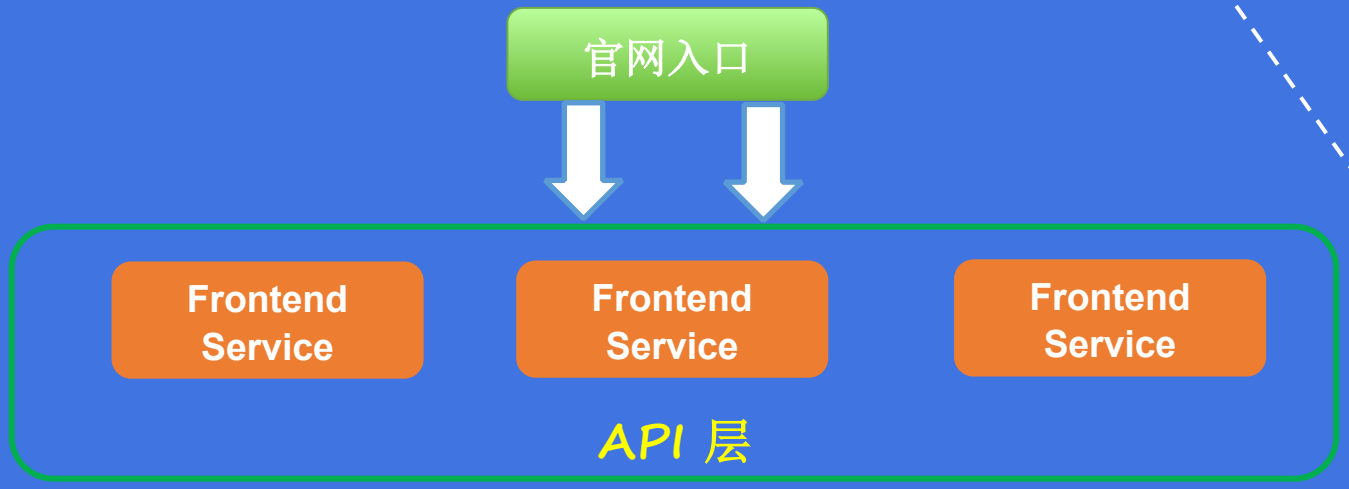
UCLOUD

案例1：.NET框架里的定时炸弹

UCLOUD

背景情况

- 对接Azure官网的前端服务(Frontend Service Fleet)
- 非高频请求（~500万请求/日)

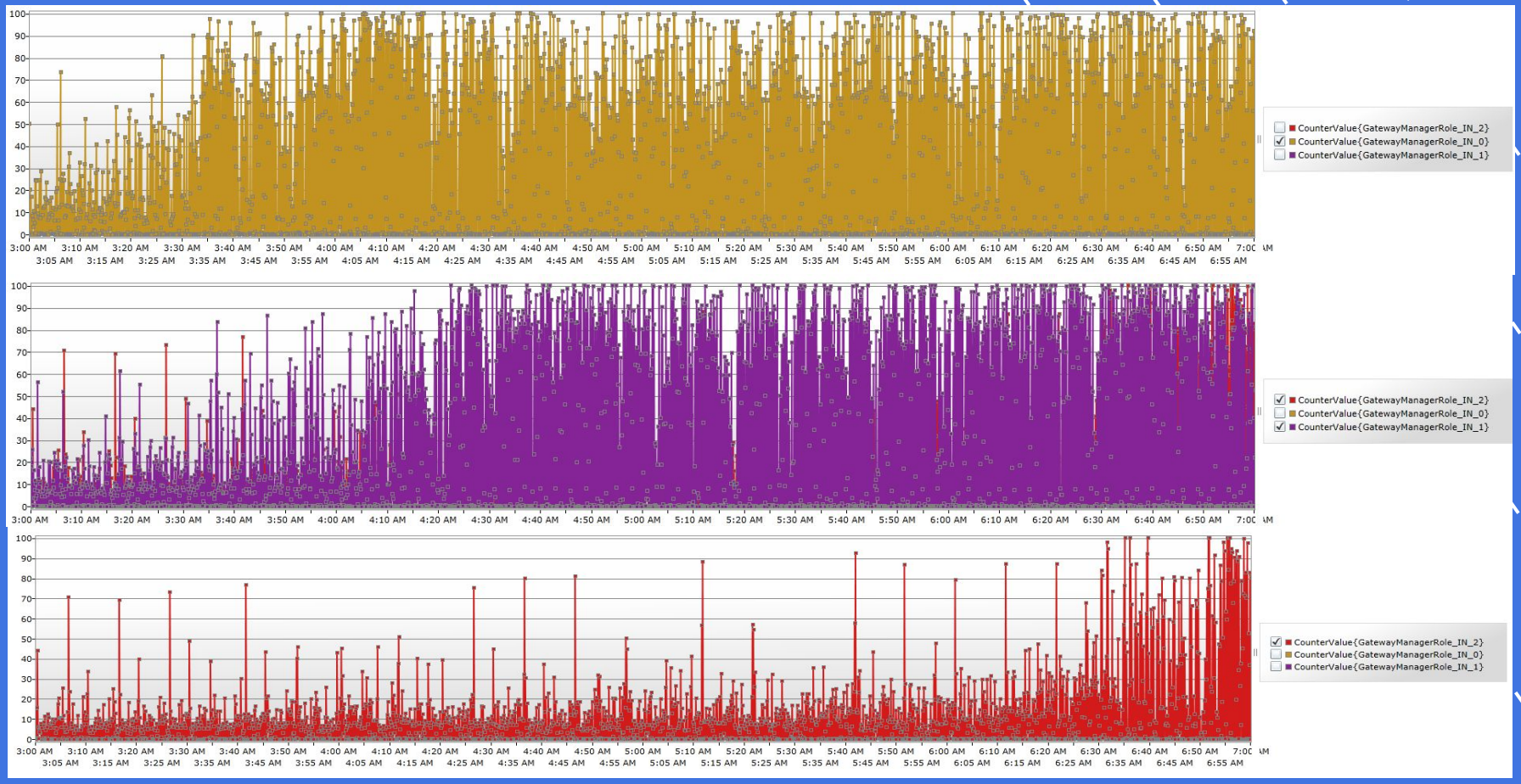


UCLOUD

事故现象

- 大约3天前发布了新的v6.0.1版frontend service
- 没有任何告警，所有监控runner状态均为绿色
- 但前端团队反应有间歇性请求超时
- 典型的“brown out”
- 请求数量没有明显增加
- 最奇怪的地方：新的v6.0.1版已经发布数天了？!

故障排查：Frontend Service的CPU Counter



UCLOUD

根因：.NET类库中的bug

- Process dump显示大量的对CPU的占用均来自于一个API: **ListClientRootCertificates**

```
X509Certificate2 x509 = new X509Certificate2()
```

- .NET类库里X509Certificates2类的构造函数里有一个临时文件句柄泄露的bug:
  - http://support.microsoft.com/kb/931908
  - 最高可用临时文件数：65000

UCLOUD

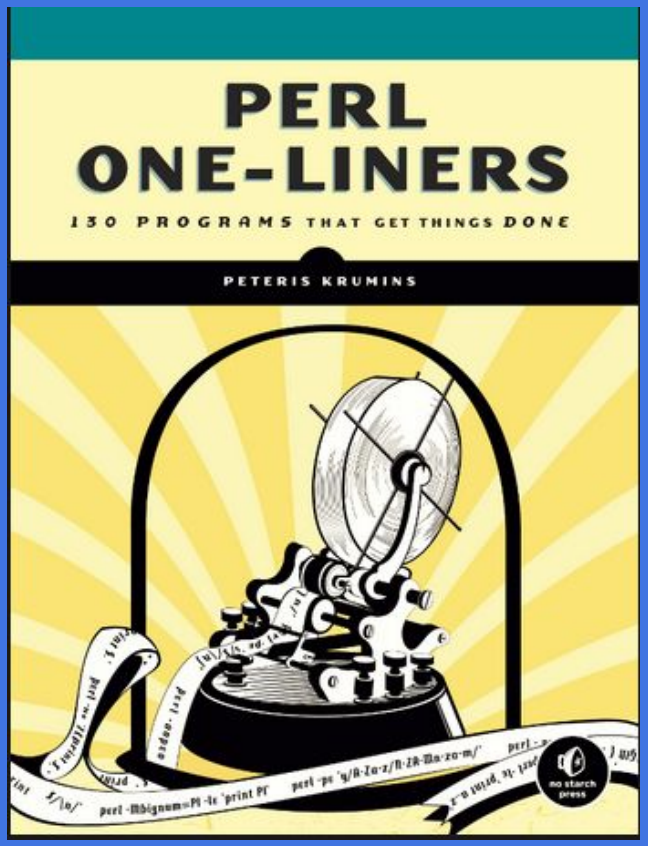
反思和教训：什么是好的监控?

- "Sign-off and Forget it" does not work. Continuous monitoring/alerting is a must.
- Features that improve monitoring, debugging and operational efficiency are as important as performance.
  - Runner only raises alert if 5 consecutive calls fail (5-minute span), not good enough
  - Design your monitoring like how you customer will use your service
- **好的监控：actionable, unambiguous, dumb**

> … In retrospect, this is more of a monitoring miss than insuffici ent test coverage. **Alerts that are non-actionable, have multip le interpretations, and are too sensitive/insensitive, are worse than no alerts.** It is inevitable to have such edge cases that are almost…

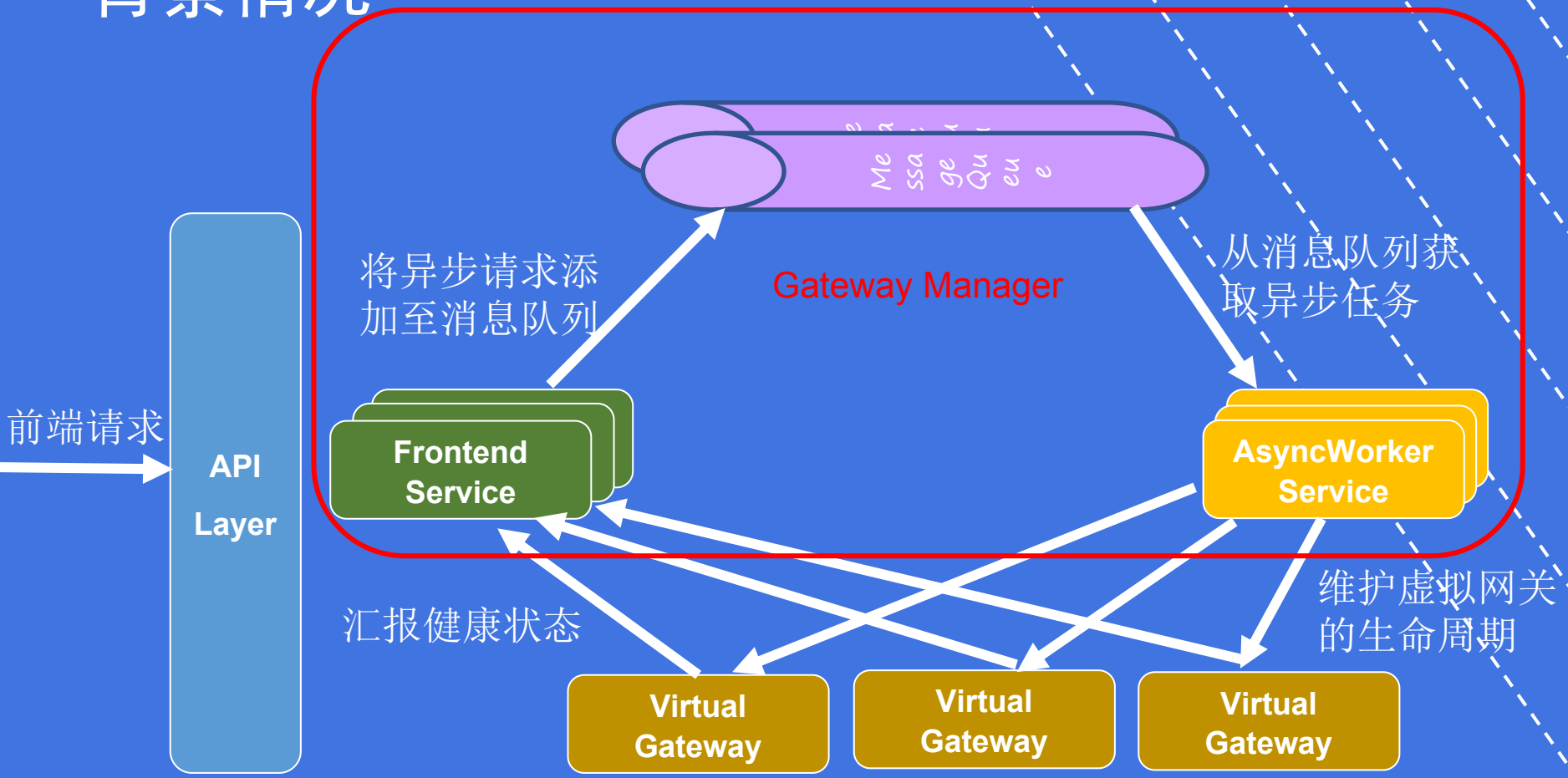
UCLOUD

案例2：一行代码的改动引发的1级事故



UCLOUD

背景情况


Message Queue
Gateway Manager
将异步请求添加至消息队列
从消息队列获取异步任务
前端请求
API Layer
Frontend Service
AsyncWorker Service
汇报健康状态
维护虚拟网关的生命周期
Virtual Gateway
Virtual Gateway
Virtual Gateway

UCLOUD

事故现象

- 服务准备发布重要的新版本
- 新版的GatewayManager部署后前端服务立刻brown-out
- 调查：花费了近一个月
- 检查了几乎所有可能的变动: WCF/MEP+/GuestOS change
- 检查了几乎所有的变量: CPU-Mem/process dump/WCF exceptions/system events/packet captures
- 尝试了多种方法: mock deployment/live deployment/simulation

ID	Title	Type	Created Date	IncidentSeverity	Changed Date	HitCounter	State	EscalationStatus	Component Team DEV	Reason	Effort
950254	Gateway Datapath useast has failed in Prod	LiveSite	08-08-2013 20:38:57	3	08-08-2013 20:48:46	3	Investigate		Rajesh Waghmare (Mindtree Consulting PVT LTD)	Unknown Issue	
950267	Gateway Datapath uscentral has failed in Prod	LiveSite	08-08-2013 20:44:22	3	08-08-2013 20:48:35	2	Investigate			Unknown Issue	
950266	Gateway Datapath asiasoutheast has failed in Prod	LiveSite	08-08-2013 20:44:10	3	08-08-2013 20:48:39	2	Investigate			Unknown Issue	
950265	Gateway Datapath useast2 has failed in Prod	LiveSite	08-08-2013 20:44:10	3	08-08-2013 20:48:35	2	Investigate			Unknown Issue	
950264	Gateway Datapath asiaeast has failed in Prod	LiveSite	08-08-2013 20:44:06	3	08-08-2013 20:48:46	2	Investigate			Unknown Issue	
950263	Gateway Datapath ussouth has failed in Prod	LiveSite	08-08-2013 20:44:06	3	08-08-2013 20:48:42	2	Investigate			Unknown Issue	
950262	Gateway Datapath europenorth has failed in Prod	LiveSite	08-08-2013 20:44:03	3	08-08-2013 20:48:35	2	Investigate		Rajesh Waghmare (Mindtree Consulting PVT LTD)	Unknown Issue	
950261	Gateway Datapath usnorth has failed in Prod	LiveSite	08-08-2013 20:44:02	3	08-08-2013 20:48:37	2	Investigate		Rajesh Waghmare (Mindtree Consulting PVT LTD)	Unknown Issue	3
950259	Gateway Deployment useast2 has failed in Prod	LiveSite	08-08-2013 20:43:59	3	08-08-2013 20:46:25	1	Investigate		Rajesh Waghmare (Mindtree Consulting PVT LTD)	Unknown Issue	2
950256	Gateway Datapath europewest has failed in Prod	LiveSite	08-08-2013 20:38:57	3	08-08-2013 20:48:47	3	Investigate		Rajesh Waghmare (Mindtree Consulting PVT LTD)	Unknown Issue	
950255	Gateway Datapath uswest has failed in Prod	LiveSite	08-08-2013 20:38:57	3	08-08-2013 20:48:43	3	Investigate			Unknown Issue	

UCLOUD

故障排查：重新审核了所有变更的代码

- public KeyToTenantMappingEntity GetEntityFromVnetId(string vnetId)
  public KeyToTenantMappingEntity GetEntityFromVnetId(Guid vnetId)

```
{
    var query = from e in TableQuery.GetBaseQuery()
            where (e.RowKey.Equals(vnetId))
            where (e.VNetId.Equals(vnetId))
            select e;

    IEnumerable<KeyToTenantMappingEntity> entities = TableQuery.RunTableQuery(query);
    return entities.FirstOrDefault();
}
```

PartitionKey	RowKey	Timestamp	AzureSubscriptionId	GatewayId	IsTenantBlacklist	TenantCertThumbprint	VNetId
0	00028b1d-7350-445b-81fa-ba493676e7ce	2013-05-14T21:13	be39470b-5b98-4f4b-l	e193654d-408e-4	False	1E64D1957E762D84F2B1	00028b1d-7350-445b-81fa-ba493676e7ce
0	000474d0-335f-4d30-8fc0-de60a14c011c	2013-05-14T21:13	6d9db675-598d-4f02-a	9dce9b09-d382-4	False	EB56F6F9D5B5DEF7210A	000474d0-335f-4d30-8fc0-de60a14c011c
0	00000000-3788-4e73-9bca-e4bf929f1bf6	2013-11-18T19:03	00000000-0000-0000-(	00000000-0000-0	False	6E6450097D10A9266F18	00000000-3788-4e73-9bca-e4bf929f1bf6
0	0006e6c4-b369-4466-9633-b7cc887a05b8	2013-05-14T21:13	e5203540-eb6e-4648-9	e5406cd5-dae5-4	False	795F26AEE71302A5F721E	0006e6c4-b369-4466-9633-b7cc887a05b8

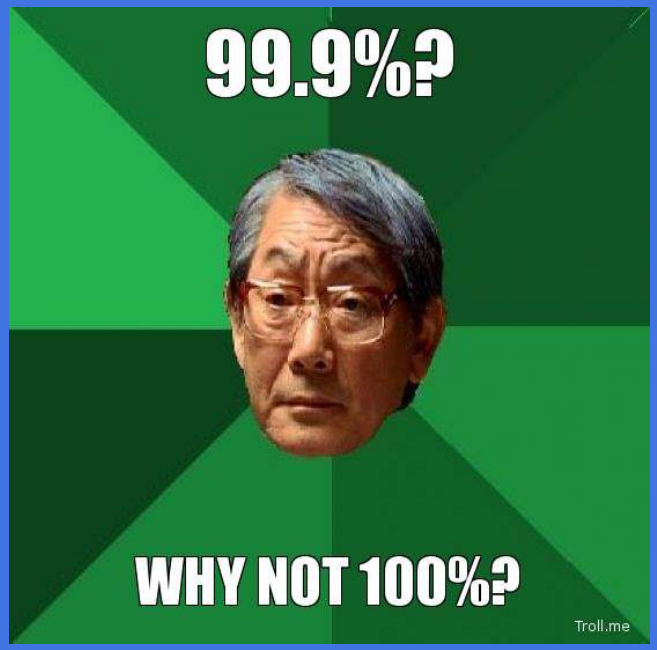
UCLOUD

反思和教训：不要盲信测试环境

- Not matter how good your test is, it is still no PRODUCTION.
- Fast rollback. Fast iteration. Incremental success
- **解耦，解耦，解耦：Partition your critical use cases.**

UCLOUD

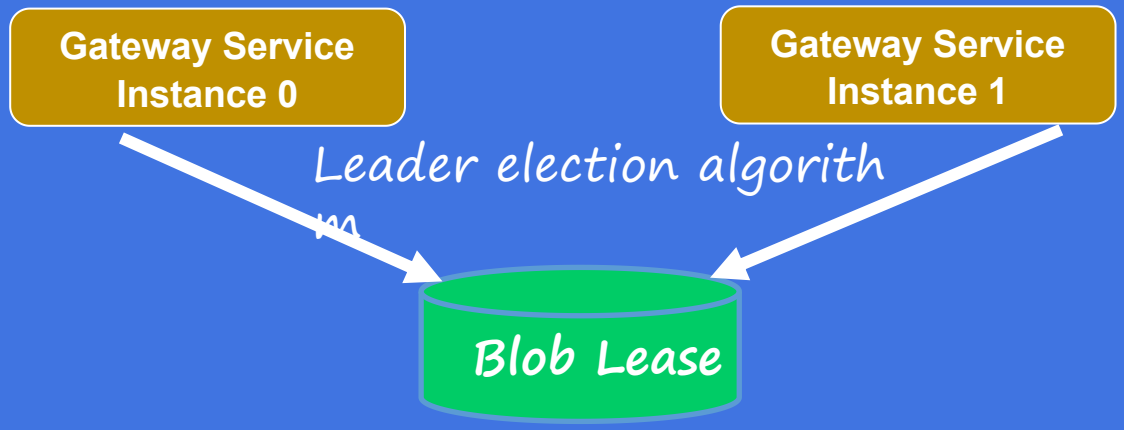
案例3：99. 99%是不够的



UCLOUD

背景情况

- USEast服务宕机
- 基于分布式锁的主从架构



UCLOUD

事故现象

- 有问题的主机都无法进行主从选举 (no primary)
- 但USEast的分布式锁服务没有问题？！

UCLOUD

故障排查

```
PublicRdfe.StorageService existingService = storageServiceList.Find((s) =>
{
    return s.ServiceName.StartsWith(storageLocationConstraint);
});
// s = "useast2xxxx" will match "storageLocationConstraint = useast"!
```

- 代码bug导致USEast的服务用了USEast2区域的分布式锁
- 但真正的问题是我们依赖于一个单点
  - 即使这个单点有99.99%的可用率
  - 即使这个单点有三份geo-distributed的replica

… One of the key design change we made in response to this flaw is to employ a 3-blob-lease leader election algorithm. **At data center level, one must have HA solution embedded into the original design in order to mitigate any potential single point of failure no matter how improbable such incident may appear to be.** If it may happen, it will happen…

UCLOUD

反思和教训：靠自己

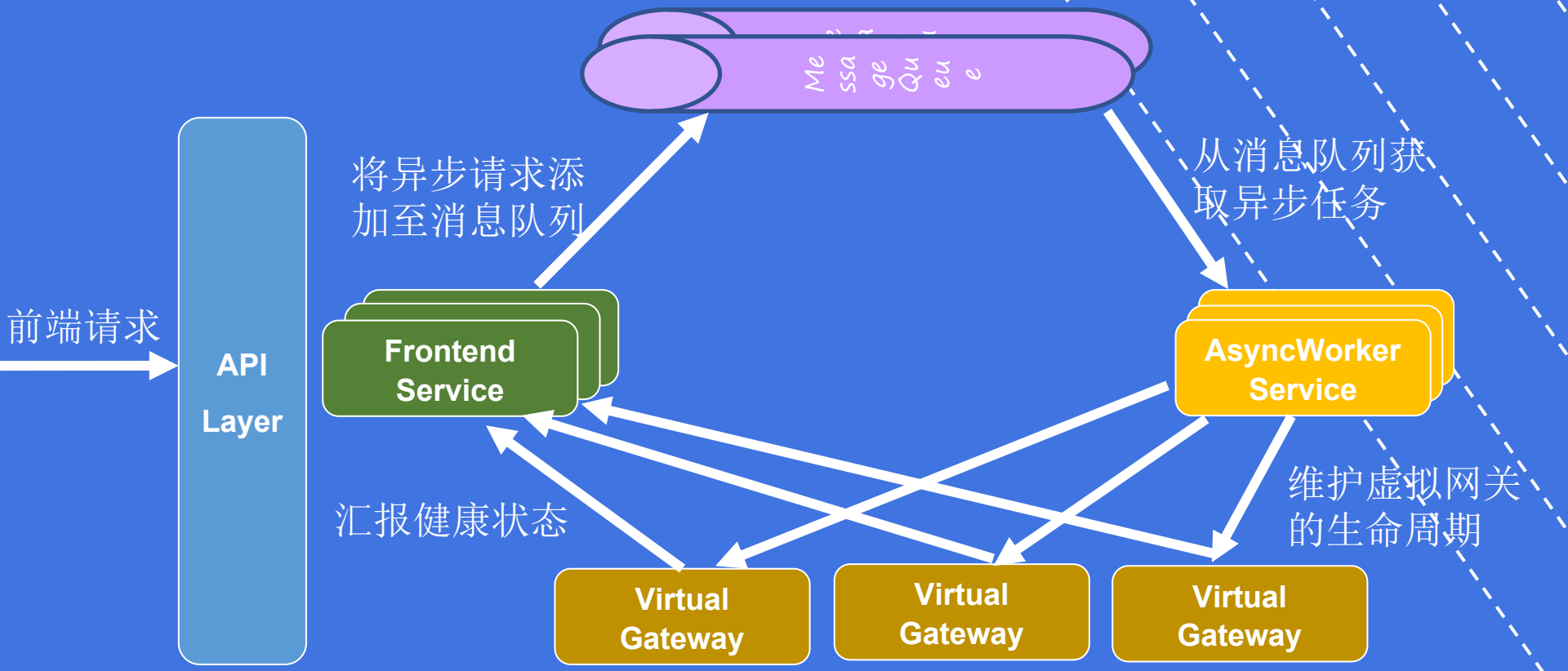
- Trust no one and rely on no one.
  - **Have a dumb overwrite safety for your key configuration**
- Is your everyday traffic representative of your worst-case traffic

UCLOUD

案例4：容错模式很重要

UCLOUD

背景情况


Message Queue
将异步请求添加至消息队列
从消息队列获取异步任务
前端请求
API Layer
Frontend Service
AsyncWorker Service
汇报健康状态
维护虚拟网关的生命周期
Virtual Gateway
Virtual Gateway
Virtual Gateway

UCLOUD

事故现象

- 控制面完全“black out”
- 增加后端服务实例只带来暂时缓解
- 偶尔有一两个异步任务可以执行
- 重启服务实例后又会慢慢进入死锁状态

ID	Title	HitCounter	Work Item Type	Assigned To	Component Team DEV	State
995026	Gateway Deployment europenorth has failed in Prod	1	RDIncident	Windows Azure Livesite	Surender Kunchala (Mindtree Consulting PVT LTD)	Investigate
995027	Gateway Deployment uscentral has failed in Prod	1	RDIncident	Windows Azure Livesite	Surender Kunchala (Mindtree Consulting PVT LTD)	Investigate
995028	Gateway Deployment usnorth has failed in Prod	1	RDIncident	Windows Azure Livesite	Surender Kunchala (Mindtree Consulting PVT LTD)	Investigate
995029	Gateway Deployment useast2 has failed in Prod	1	RDIncident	Windows Azure Livesite	Surender Kunchala (Mindtree Consulting PVT LTD)	Investigate
995030	Gateway Deployment uswest has failed in Prod	1	RDIncident	Windows Azure Livesite	Surender Kunchala (Mindtree Consulting PVT LTD)	Investigate
995031	Gateway Deployment useast has failed in Prod	1	RDIncident	Windows Azure Livesite	Surender Kunchala (Mindtree Consulting PVT LTD)	Investigate
995033	Gateway Deployment europewest has failed in Prod	1	RDIncident	Windows Azure Livesite	Surender Kunchala (Mindtree Consulting PVT LTD)	Investigate
995034	Gateway Deployment asiaeast has failed in Prod	1	RDIncident	Windows Azure Livesite	Surender Kunchala (Mindtree Consulting PVT LTD)	Investigate
995035	Gateway Deployment europewest has failed in Prod	1	RDIncident	Windows Azure Livesite	Surender Kunchala (Mindtree Consulting PVT LTD)	Investigate

UCLOUD

故障排查

```
var serverCertFile = dir + "\\" + parameters.VnetName + CertExtension;
int exitCode = PlatformUtils.StartProcess(iexpress, "/N "
        + TextUtils.EvaluateTemplatizedString(SedFileName, serverCertFile),
        out output, out error, null, true, dir);
```

- 前端改动了验证逻辑，允许了更大的字符集
  - iexpress.exe /N abc.cer
  - iexpress.exe /N a bc.cer
- 每个去执行有问题的异步任务的后端线程都会死锁
- 每4分钟损失一个线程，N*4分钟后，整个系统black out

UCLOUD

反思和教训：好的容错模式

- Sanitize user input (durrr).
- Watch out for locking of critical resources.
- Does your failure mode scale?
- Do not brown out. Always fail fast.
- Dig into unusual outages or behavior, even if brief.

UCLOUD

总结

- Continuous monitoring/alerting is a must.
- Features that improve monitoring, debugging and operational efficiency are as important as performance.
- Not matter how good your test is, it is still no PRODUCTION.
- Fast rollback. Fast iteration. Incremental success.
- Partition your critical use cases.
- Have a dumb overwrite safety for your key configuration.
- Is your everyday traffic representative of worst-case traffic?
- Watch out for locking of critical resources.
- Does your failure mode scale?
- Do not brown out. Always fail fast.
- Dig into unusual outages or behavior, even if brief.

UCLOUD

加入我们：UCloud公有云运维架构师

UCLOUD

加入我们：UCloud公有云运维架构师

- **工作地点**：上海杨浦
- **工作职责**：负责客户公司的故障定位、服务恢复、性能优化、架构设计等；负责云平台新技术的研究
- **职位要求**：
- 精通Linux操作系统
- 精通TCP/IP协议栈；熟悉至少一种TCP/HTTP应用体系（Nginx/Apache/HaProxy/LVS等）
- 熟悉至少一种虚拟化技术（vSphere/Xen/KVM/LXC等）
- 熟悉常见应用系统架构（电商/游戏/社交/视频等）
- **联系方式**： 微信13501916328