



中國計算機學會
CHINA COMPUTER FEDERATION



Nightingale



GitLink
— 确实 · 开源 —

第二届CCF·夜莺开发者创新论坛

中国北京 2024.7.26

主办方: 中国计算机学会 | 承办方: CCF开源发展委员会、夜莺项目开源社区



中國計算機學會
CHINA COMPUTER FEDERATION



Nightingale



GitLink
— 确实 · 开源 —

可观测系统的多机房实践

小米 马千里

中国北京 2024.7.26

主办方: 中国计算机学会 | 承办方: CCF开源发展委员会、夜莺项目开源社区

大纲

能力现状

架构介绍

关键问题

当前工作

能力现状

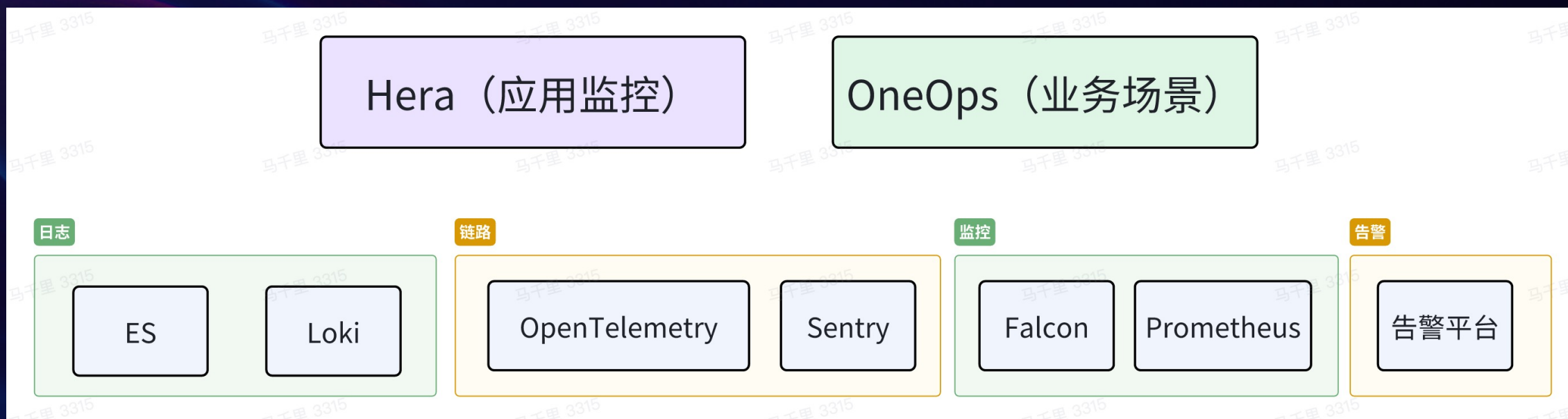
中国北京 2024.7.26



中國計算機學會
CHINA COMPUTER FEDERATION



能力现状



架构介绍

中国北京 2024.7.26

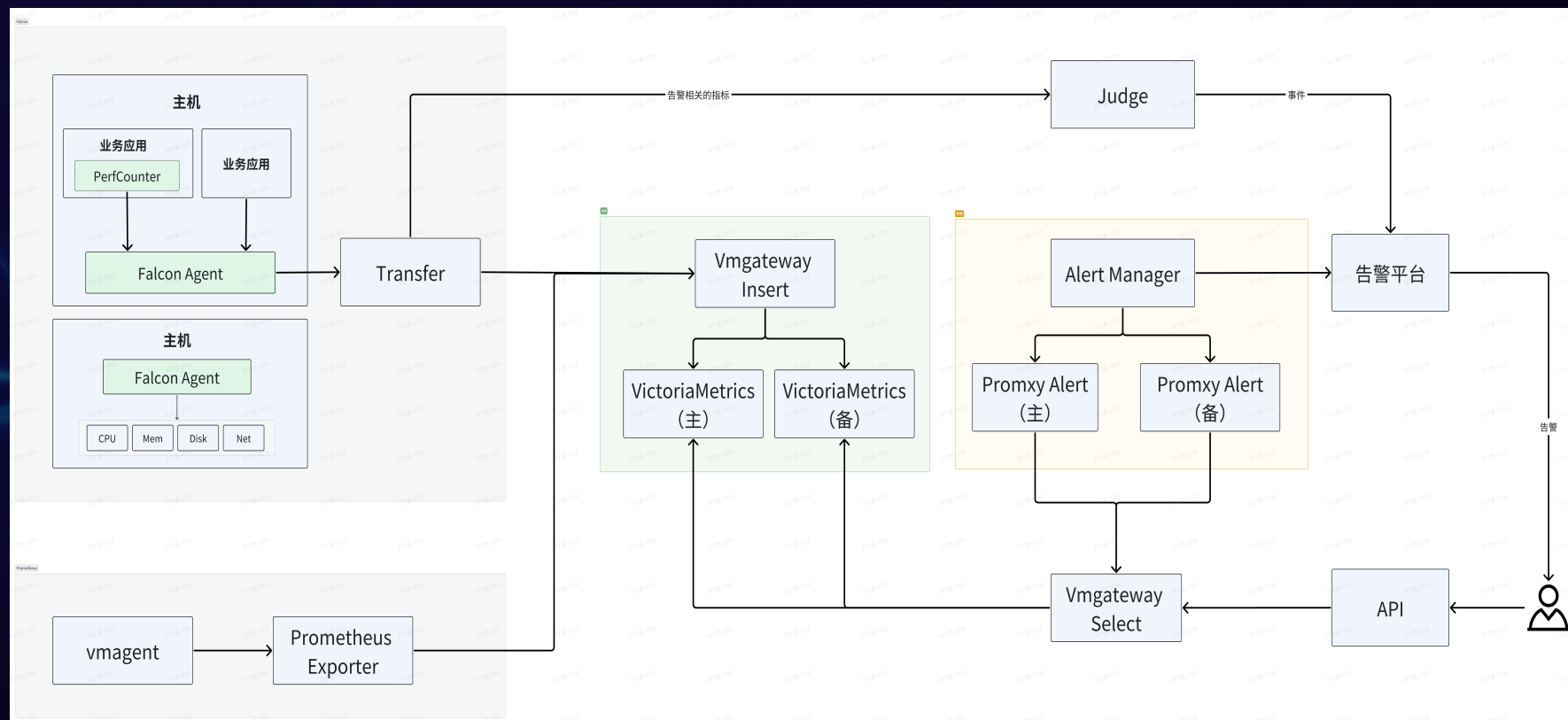


中國計算機學會
CHINA COMPUTER FEDERATION



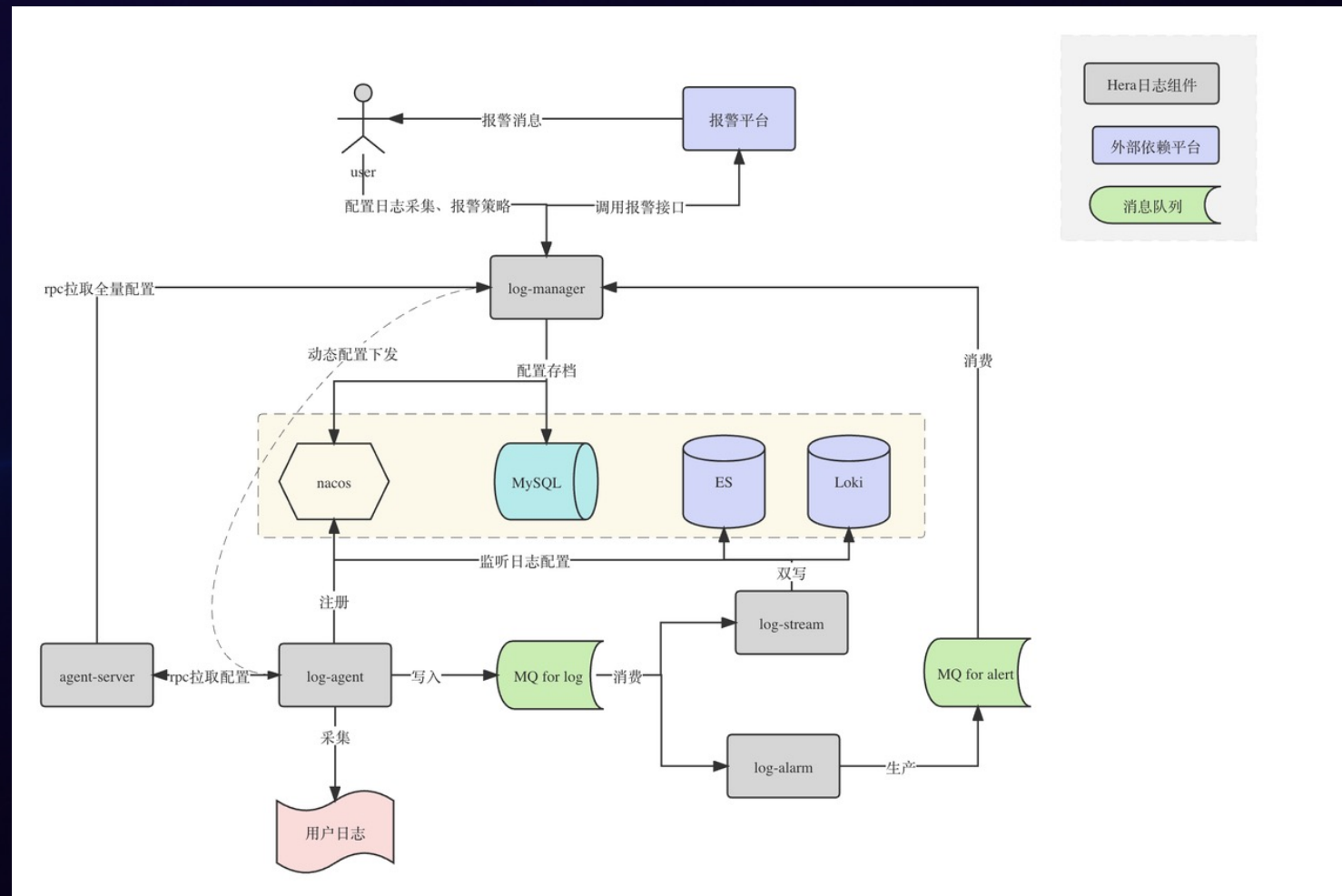
监控架构

- 还保留有 Open-Falcon 的元素，比如 Falcon Agent、Transfer、Judge。
- 增加了 Prometheus 相关的能力。
- 将 Open-Falcon 的 Alarm 提取为告警平台。
- VMgatewayInsert 和 VMgatewaySelect 为自研组件，起到代理和主备切换的作用。
- VM 在各机房本地化部署，双 AZ 部署。



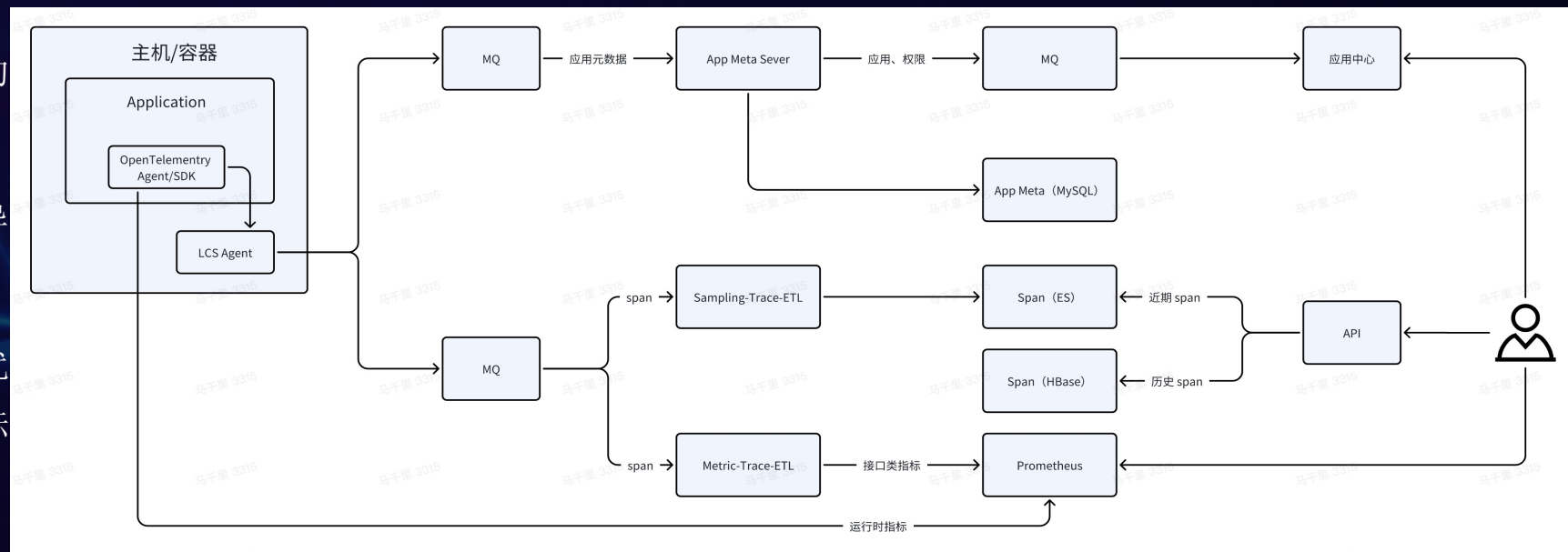
日志架构

- 底层以 ES 为主要方案。
- Loki 用作冷备存储，同时在多个场景有深度使用。
- ES、Loki 在各机房本地化部署。
- 告警接入告警平台。



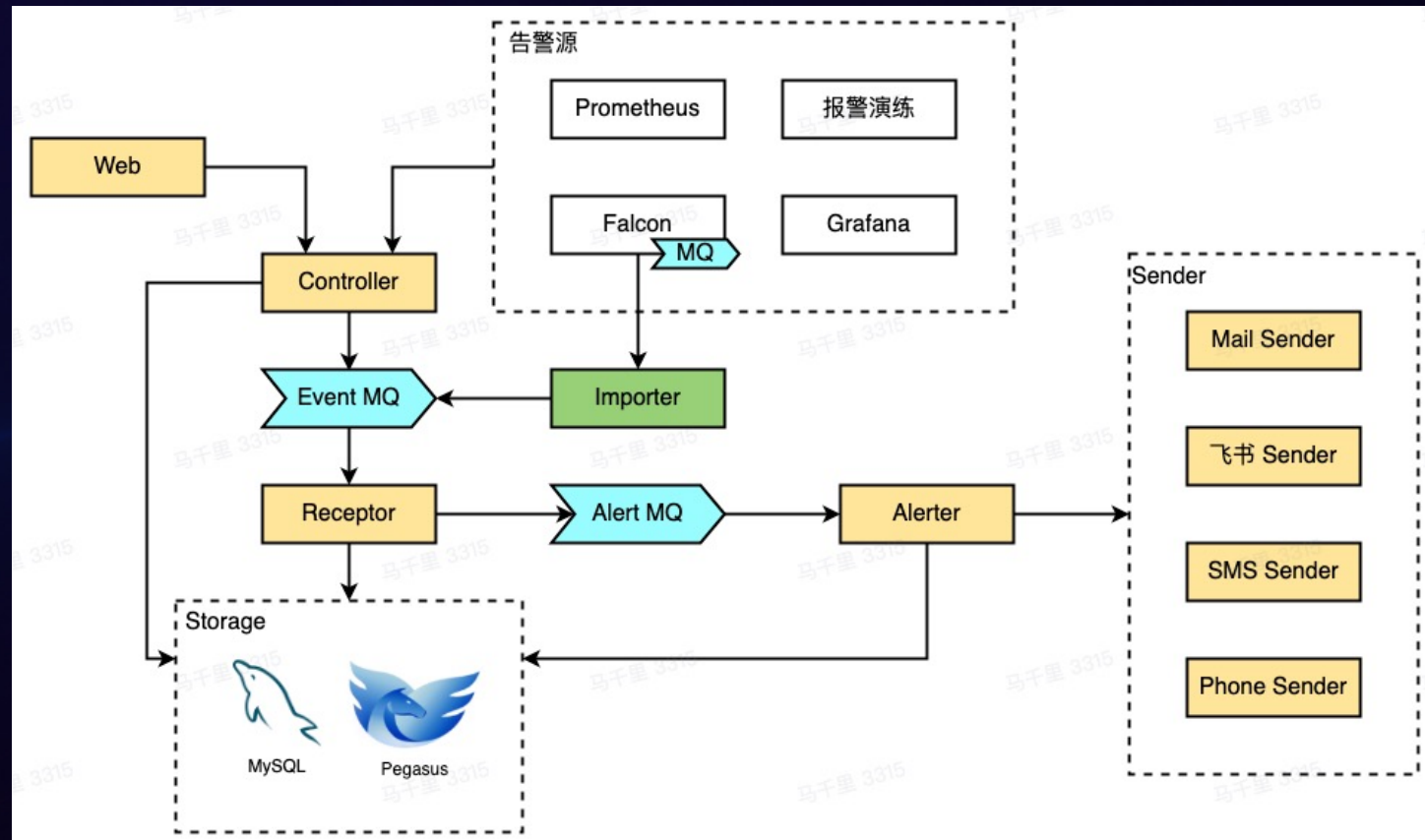
链路架构

- 应用：“应用”来自不同的平台，包括主机、容器平台、Flink 等。
- 应用中心：在应用元数据基础上，提供的应用的管理能力。
- Sampling-Trace-ETL：尾采样。只保留异常、慢查相关的 trace。性价比高。
- Metric-Trace-ETL：服务端计算指标。优点是和客户端语言解耦，不足是会在指标产生额外的维度，有高基数的风险。
- 底层以 ES 为主要方案，在各机房本地化部署。



告警架构

- 从 Open-Falcon 的 Alarm 演化而来。增加了“告警源”，以及相关样式、策略的抽象。
- 告警服务只在内部署，各机房的告警事件统一回国发送。
- 为了应对专线故障，开通了公网的加密链路，作为备用通道。
- 设计之初，参考了 PagerDuty、FlashDuty 等产品。致敬。



关键问题

中国北京 2024.7.26



中國計算機學會
CHINA COMPUTER FEDERATION



问题一：大文件 merge 导致查询超时

简介

- 现象：查图功能异常。
- 原因：VM 的文件合并没有被有效限制，大文件 merge 导致严重写放大，单个节点负载升高，集群的查询请求被间接 delay 并超时。

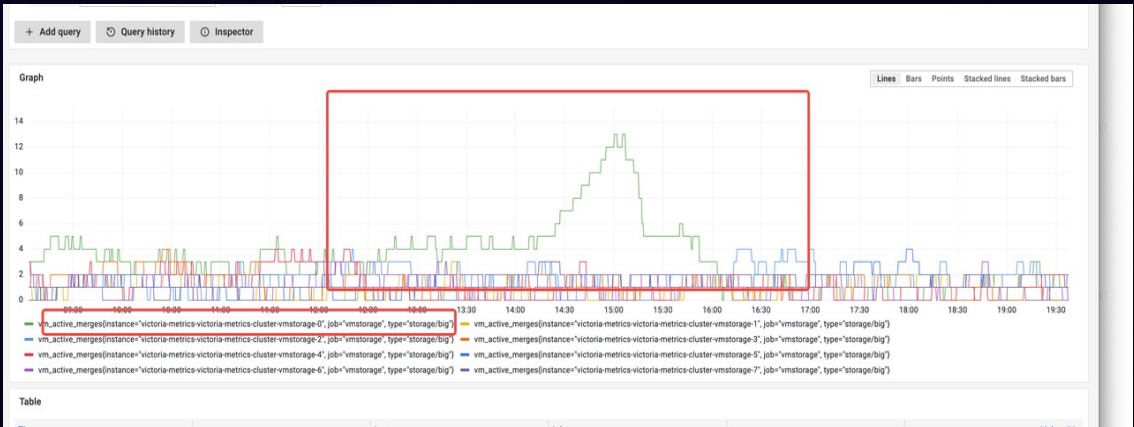
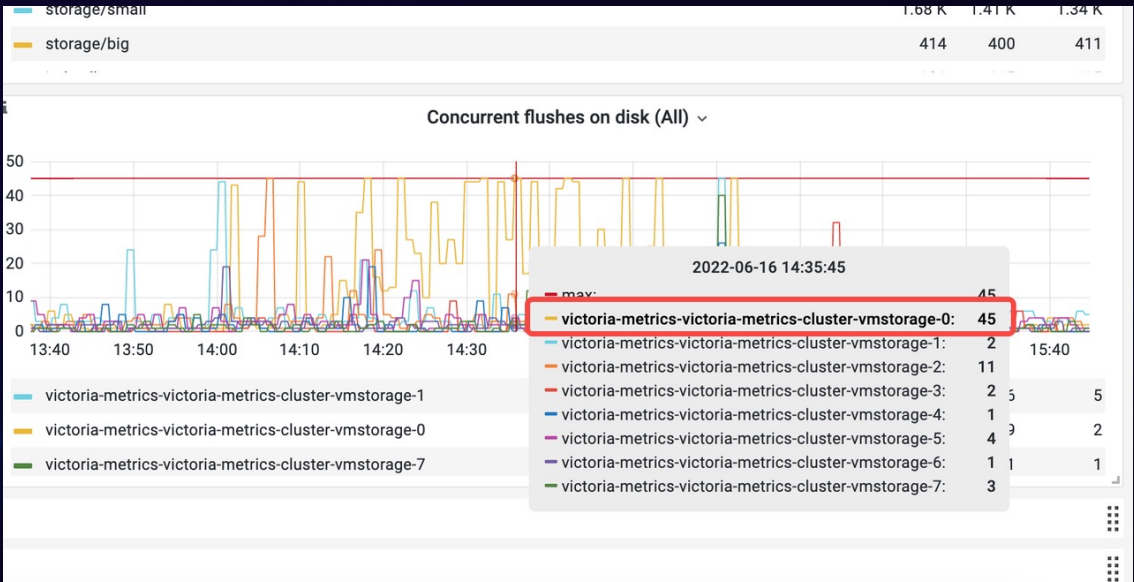
方案

- 故障期间将查询切换到备用集群
- 更新 VM 版本
- 增加告警，比如 vm_activated_merge 告警

问题一：大文件 merge 导致查询超时

排查

- 1. 单个节点的刷盘线程数达到 CPU 核数限制（45 核）；该节点请求延迟严重，锁定故障节点。
- VM 查询需要等到所有节点返回结果，单个节点故障会导致整个集群查询超时。
- 2. vmselect 到 vmstorage 之间 RPC 长连接，不需要重复建立连接，但在故障期间有大量建连。
- “cannot process vmselect conn xx.xx.xx.xx:35630: cannot process vmselect request: cannot read rpcName: cannot read data size: cannot read data in 8.730 seconds: read tcp4 xx.xx.xx.xx:8401->xx.xx.xx.xx:35630: read: connection reset by peer”
- 3. 查询大量积压。VM 优先保证写入，写入延迟高时会自动 delay 查询。
- 4. 当天故障节点的合并文件数在故障期间飙升，可以确定是引起大量磁盘 IO 的原因（近一周内：故障时间外合并大文件数量不超过 4 个，故障期间达到 12 个）
- 5. 社区 2022 年 06 月 07 日修复了该问题（[PR2673](#)）。



问题二：存储集群容量不足，迁移数据时影响写入

简介

- 现象：监控数据有丢失，影响监控、报警功能。
- 原因：存储集群迁移过快，存储集群容量不足，一个节点发生故障导致整个集群写入异常。

方案

- 适当扩容，确保集群能承载当前 150% 左右的写入流量。
- 调整 vmagent 组件参数，对写入流量的上限进行限制。
- 其他全部 vmagent 增加数据积压的监控报警，采集端出现积压时能够第一时间感知。

问题二：存储集群容量不足，迁移数据时影响写入

排查

1. 迁移集群时，为了加快迁移进度，迁移过程中让一个节点承载了线上全部写入流量。该节点触发了 compaction 过程，负载升高，数据无法写入该节点。
2. 存储集群发生上述故障后，采集组件 **vmagent** 无法把采集到的监控指标数据全部写入存储集群，未写入数据会保存在磁盘上，随后进行无限重试，即发生数据积压。
3. 在数据积压发生后，**vmagent** 会尝试同时把积压的数据和新采集的数据写入到存储集群，这使得存储集群接收到的该组件的写入数据越来越多，造成“雪崩”现象。

问题三：集群负载高，单节点故障影响整个集群

简介

现象：监控数据有丢失，影响监控、报警功能。

原因：存储集各个节点的内存使用率太高，导致一个节点故障，其它节点再接收它的写入流量后陆续发生OOM，导致集群长时间不可用。

方案

- 故障期间将查询切换到备用集群
- 适当扩容，使集群保持正常水位，能够容忍单点故障
- 增加集群查询 QPS 大幅下降、查询响应时间大幅上升等情况配置 P0 报警

It is recommended leaving the following amounts of spare resources:

- 50% of free RAM across all the node types for reducing the probability of OOM (out of **memory**) crashes and slowdowns during temporary spikes in workload.
- 50% of spare CPU across all the node types for reducing the probability of slowdowns during temporary spikes in workload.
- At least 20% of free storage space at the directory pointed by `-storageDataPath` command-line flag at `vmstorage` nodes. See also `-storage.minFreeDiskSpaceBytes` command-line flag [description for vmstorage](#).

问题三：集群负载高，单节点故障影响整个集群

排查

1. 所有节点都发生重启。
2. 一个节点重启原因是 Error，其它节点重启原因是 OOMKilled。
3. 从重启顺序上初步定位故障节点。
4. 检测发现故障节点内存异常。故障节点服务也有相应错误日志。



```
[{"log": "\ts": "2024-06-01T18:00:05.336Z", "level": "info", "caller": "VictoriaMetrics/lib/vmselectapi/server.go:192", "msg": "processing vm", "time": "2024-06-01T18:00:05.336Z"}, {"log": "unexpected fault address 0x7cb99d553ab3\n", "stream": "stderr", "time": "2024-06-01T18:00:05.445015862Z"}, {"log": "fatal error: panic\n", "stream": "stderr", "time": "2024-06-01T18:00:05.445015862Z"}, {"log": "signal SIGBUS: bus error code=0x4 addr=0x7cb99d553ab3 pc=0x6cbd17\n", "stream": "stderr", "time": "2024-06-01T18:00:05.448291062Z"}, {"log": "\n", "stream": "stderr", "time": "2024-06-01T18:00:05.448309034Z"}, {"log": "goroutine 3869297712 [running]:\n", "stream": "stderr", "time": "2024-06-01T18:00:05.448312105Z"}, {"log": "runtime.throw(0x8ce5822, 0x41284b?)\n", "stream": "stderr", "time": "2024-06-01T18:00:05.483036295Z"}, {"log": "\u0009runtime.panic.go:1047 +0x5d fp=0xc25eaab1b8 sp=0xc25eaab188 pc=0x43815d\n", "stream": "stderr", "time": "2024-06-01T18:00:05.483049655Z"}, {"log": "runtime.sigpanic(0\n", "stream": "stderr", "time": "2024-06-01T18:00:05.483053144Z"}, {"log": "\u0009runtime.signal_unix.go:832 +0x125 fp=0xc25eaab208 sp=0xc25eaab1b8 pc=0x44e765\n", "stream": "stderr", "time": "2024-06-01T18:00:05.4830557C"}, {"log": "github.com/VictoriaMetrics/fastcache.(*bucket).Set(0xc050017d20, {0xc17bf4b400, 0x200?, 0x6ffaa4373e76f31e?}, {0xc650f134a0, 0x85f160?, 0xc2", "stream": "stderr", "time": "2024-06-01T18:00:05.483067862Z"}, {"log": "\u0009github.com/VictoriaMetrics/fastcache@v1.10.0/fastcache.go:351 +0x297 fp=0xc25eaab330 sp=0xc25eaab208 pc=0x6cbd17\n", "stream": "stderr", "time": "2024-06-01T18:00:05.483070432Z"}, {"log": "github.com/VictoriaMetrics/fastcache.(*Cache).Set(0xc050010000, {0xc17bf4b400, 0x69, 0x200}, {0xc650f134a0, 0x10, 0x10})\n", "stream": "stderr", "time": "2024-06-01T18:00:05.483085842Z"}, {"log": "\u0009github.com/VictoriaMetrics/fastcache@v1.10.0/fastcache.go:151 +0xa5 fp=0xc25eaab380 sp=0xc25eaab330 pc=0x6cb345\n", "stream": "stderr", "time": "2024-06-01T18:00:05.483101252Z"}, {"log": "github.com/VictoriaMetrics/fastcache.(*Cache).SetBig(0xc050010000, {0xc17bf4b400, 0x69, 0x200}, {0xc0bc9f4000, 0x11, 0x4000})\n", "stream": "s", "time": "2024-06-01T18:00:05.483110672Z"}, {"log": "\u0009github.com/VictoriaMetrics/fastcache@v1.10.0/bigcache.go:64 +0x485 fp=0xc25eaab418 sp=0xc25eaab380 pc=0x6ca765\n", "stream": "stderr", "time": "2024-06-01T18:00:05.483125238Z"}, {"log": "github.com/VictoriaMetrics/VictoriaMetrics/lib/workingsetcache.(*Cache).SetBig(0xc0bc9f4000?, {0xc17bf4b400?, 0xc5678780a8?, 0xc25eaab4f0?}, {0xc0bc9f4000?, 0xc07fda", "stream": "stderr", "time": "2024-06-01T18:00:05.483152389Z"}, {"log": "\u0009github.com/VictoriaMetrics/VictoriaMetrics/lib/workingsetcache/cache.go:376 +0x7b fp=0xc25eaab470 sp=0xc25eaab418 pc=0x6d255b\n", "stream": "stderr", "time": "2024-06-01T18:00:05.483164333Z"}, {"log": "github.com/VictoriaMetrics/VictoriaMetrics/lib/storage.(*indexDB).putToTagFiltersCache(0xc29077c240, 0xc630c0?, {0x0, 0x0, 0x0}, {0xc17bf4b400, 0x69, 0x200})\n", "stream": "s", "time": "2024-06-01T18:00:05.483178842Z"}]
```

kube/config-c3-prod grep victoria-metrics-victoria-metrics-cluster-vmstorage				
-vmstorage-0	1/1	Running	3	404d
-vmstorage-1	1/1	Running	3	404d
-vmstorage-10	1/1	Running	1	219d
-vmstorage-11	1/1	Running	1	219d
-vmstorage-2	1/1	Running	2	404d
-vmstorage-3	1/1	Running	3	404d
-vmstorage-4	1/1	Running	1	297d
-vmstorage-5	1/1	Running	2	404d
-vmstorage-6	1/1	Running	2	404d
-vmstorage-7	1/1	Running	2	386d
-vmstorage-8	1/1	Running	3	382d
-vmstorage-9	1/1	Running	2	333d

当前工作

中国北京 2024.7.26

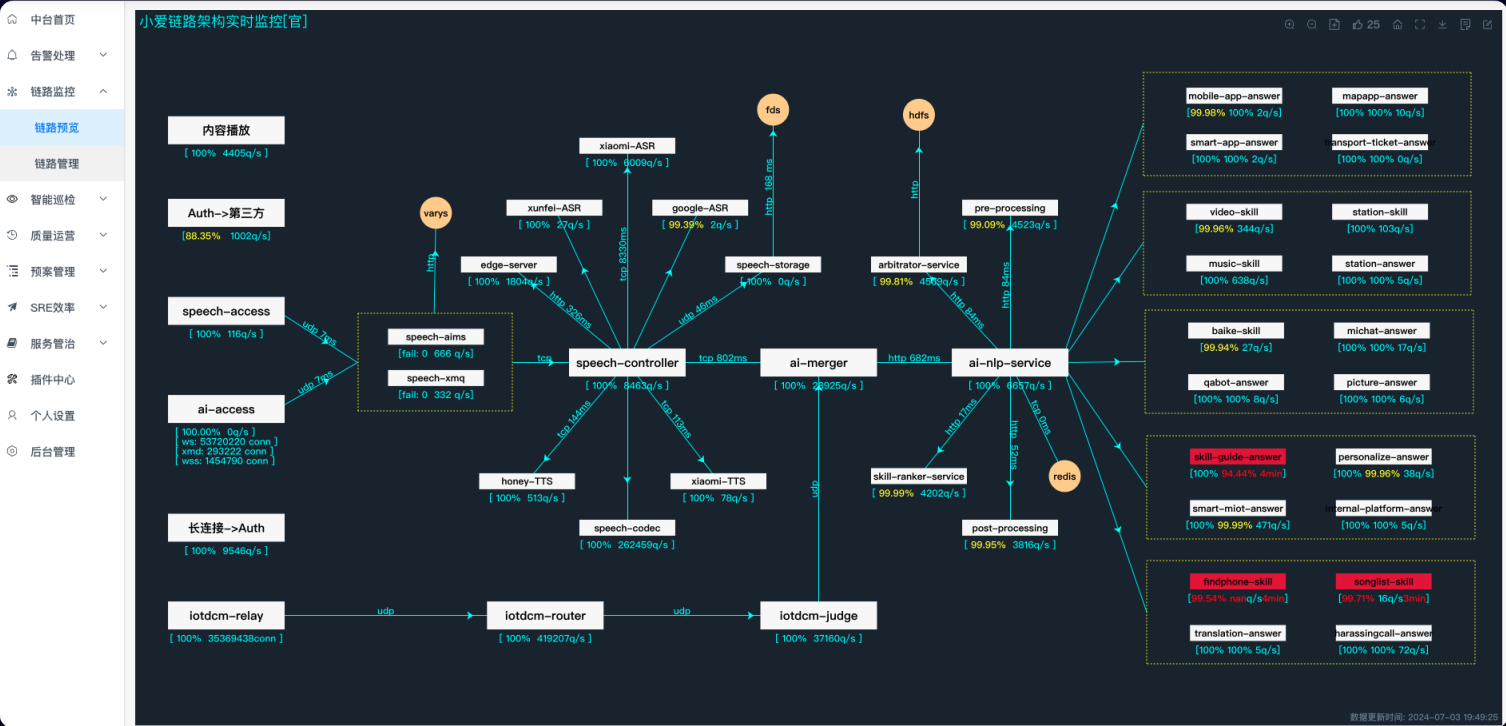
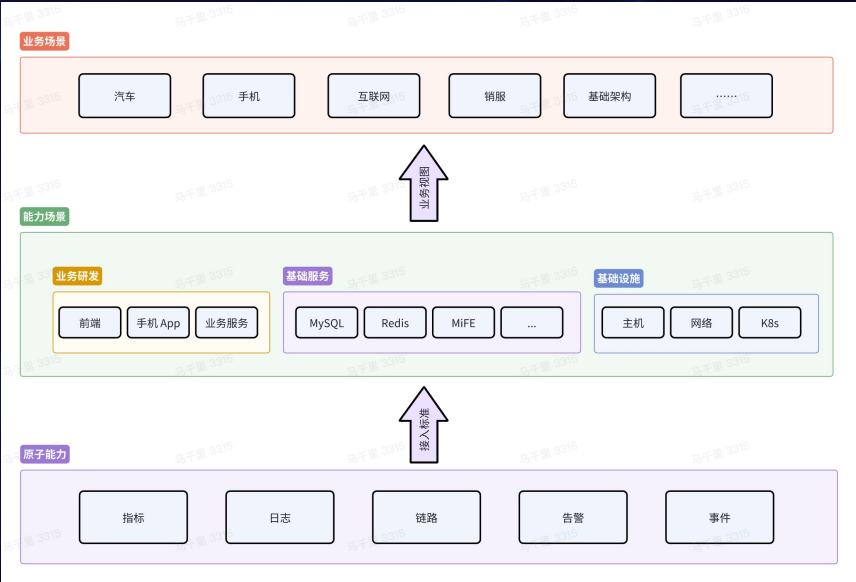


中國計算機學會
CHINA COMPUTER FEDERATION



TODO

- 1. 沉淀业务场景、北极星指标
- 2. 优化原子能力体验
- 3. 数据治理
- 4. 自动化运维



感谢大家的时间
Thank you for listening