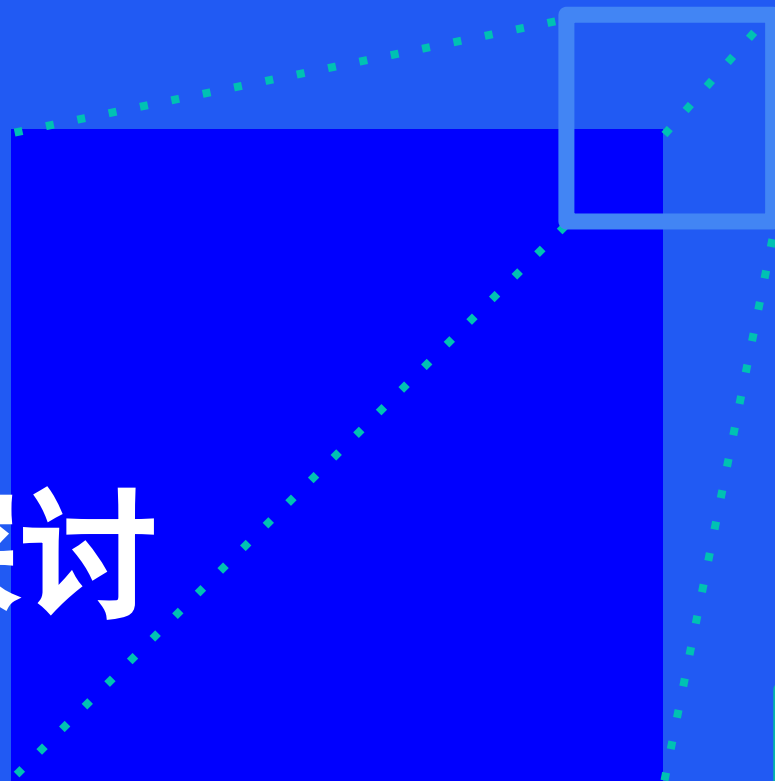


Elasticsearch

异地容灾建设方案探讨



关于分享者

曾 勇

极限科技创始人兼 CEO

Medcl, 前 Elastic 中国第一位员工 (全球 ~200 号), 12 年+ Elasticsearch 使用经验, Elastic Principal Consulting Architect, Elastic 6 年原厂工作经验, 前 Elastic 亚太区布道师, 前 Elastic 中国区咨询业务负责人, 前 Elastic 官方培训讲师, Elastic 中文社区发起人、主席, 《Elasticsearch 搜索开发实战》作者, 《Elasticsearch 权威指南》中文译版总编, Elasticsearch 若干开源插件工具作者, 阿里云 MVP, 关注高并发、分布式、搜索、中文自然语言处理等。





今天的主题

Elasticsearch 异地容灾



Why 异地容灾?

Elasticsearch 不是分布式的么？高可用？主副本？快照？备份？

Why 异地容灾?



Elasticsearch 已成为
企业主要 Infra



容灾建设即
保障业务连续型



在线后备系统
随时可切换



Elasticsearch 容灾设计要点

Elasticsearch 作为一个流行的分布式搜索和分析系统，本身提供了主副本冗余机制，并且也支持通过快照来进行周期备份，但是对于关键的业务场景，可能还需要考虑数据中心级别的异地多活备份。

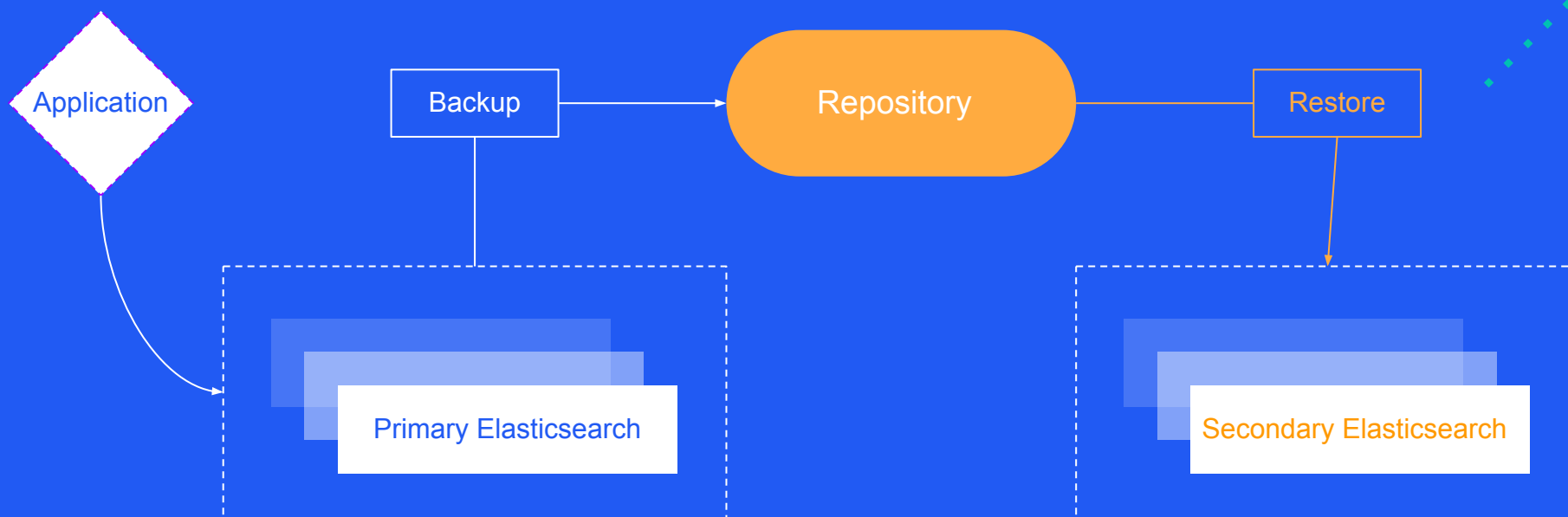
一致性	时效性	顺序性
可验证	可恢复	灵活性



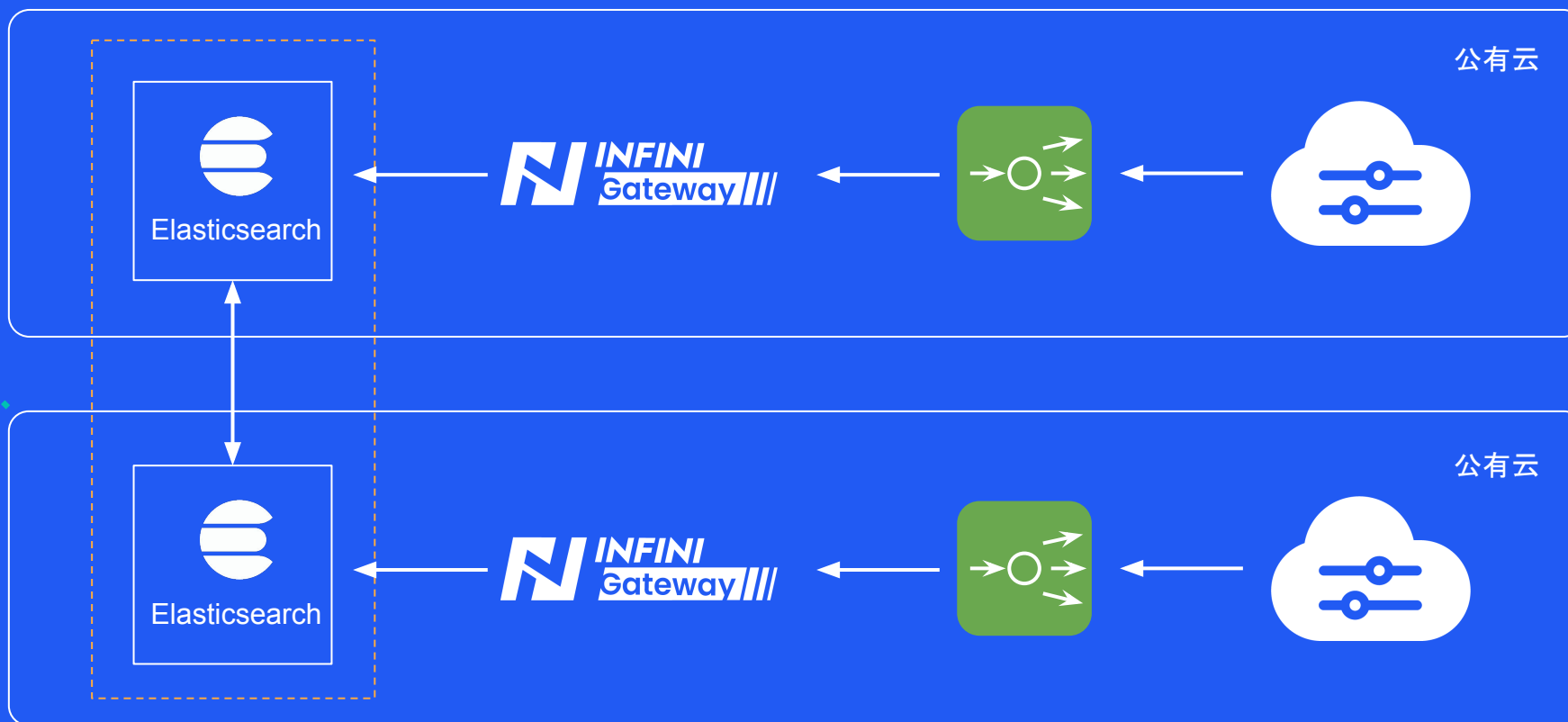
Elasticsearch

常见容灾方案介绍

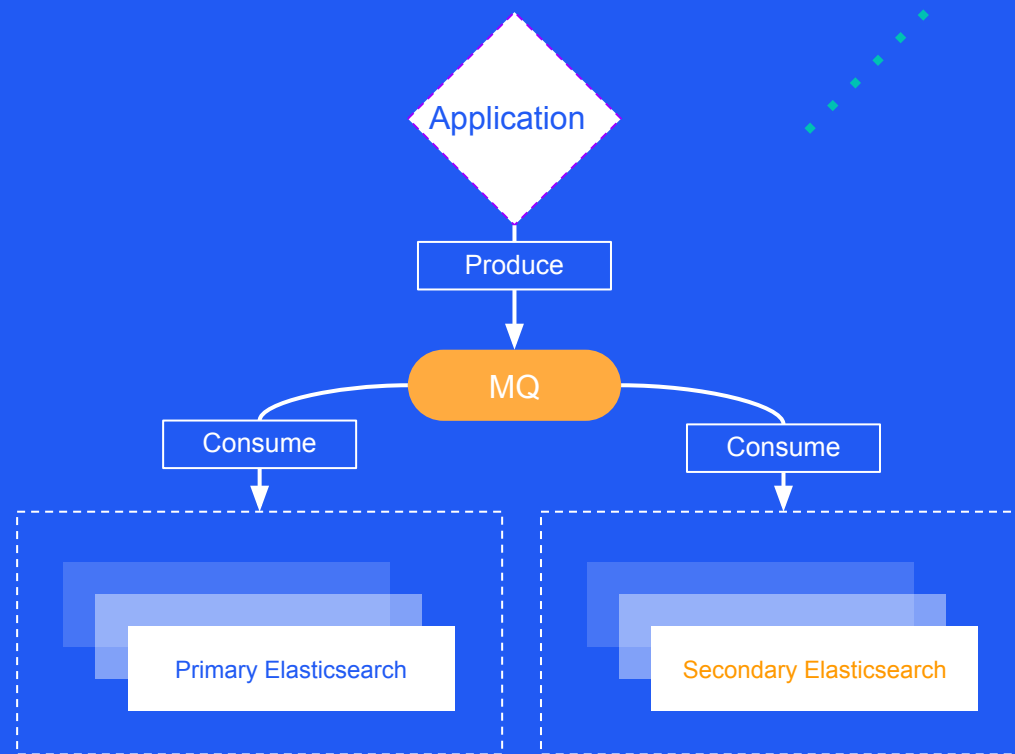
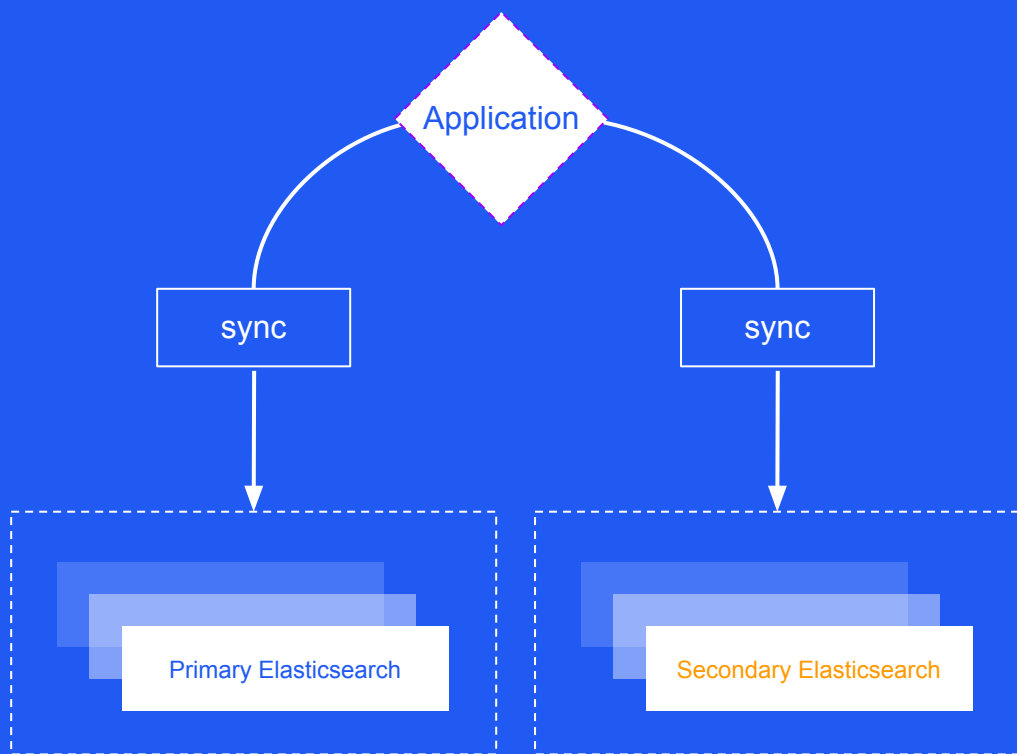
定期快照-增量备份/还原



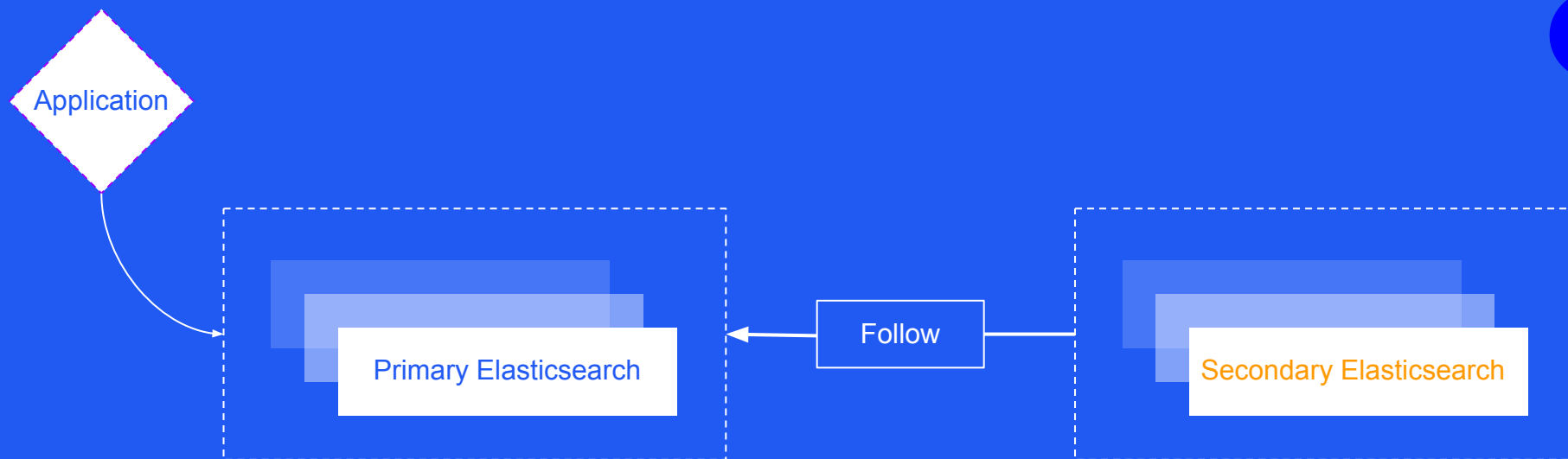
跨 Zone 集群



双写-应用/MQ双写



主从复制 单向订阅





还有其他选择么？



Primary
Elasticsearch

Secondary
Elasticsearch



什么是 INFINI Gateway?



面向 Elasticsearch 的 高性能应用网关

极限网关 (INFINI Gateway) 是一个面向 Elasticsearch 的高性能应用网关，它包含丰富的特性，使用起来也非常简单。

极限网关工作的方式和普通的反向代理一样，我们一般是将网关部署在 Elasticsearch 集群前面，将以往直接发送给 Elasticsearch 的请求都发送给网关，再由网关转发给请求到后端的 Elasticsearch 集群。

因为网关位于在用户端和后端 Elasticsearch 之间，所以网关在中间可以做非常多的事情，比如可以实现索引级别的限速限流、常见查询的缓存加速、查询请求的审计、查询结果的动态修改等等。



INFINI Gateway 特点

“极限网关”最懂 Elasticsearch，其在设计的时候就综合考虑了很多和 Elasticsearch 相关的业务场景及特点，基于此打造了很多完美契合 Elasticsearch 的众多实用功能



轻量级



极致性能



跨版本支持



可观测性



高可用



灵活可扩展

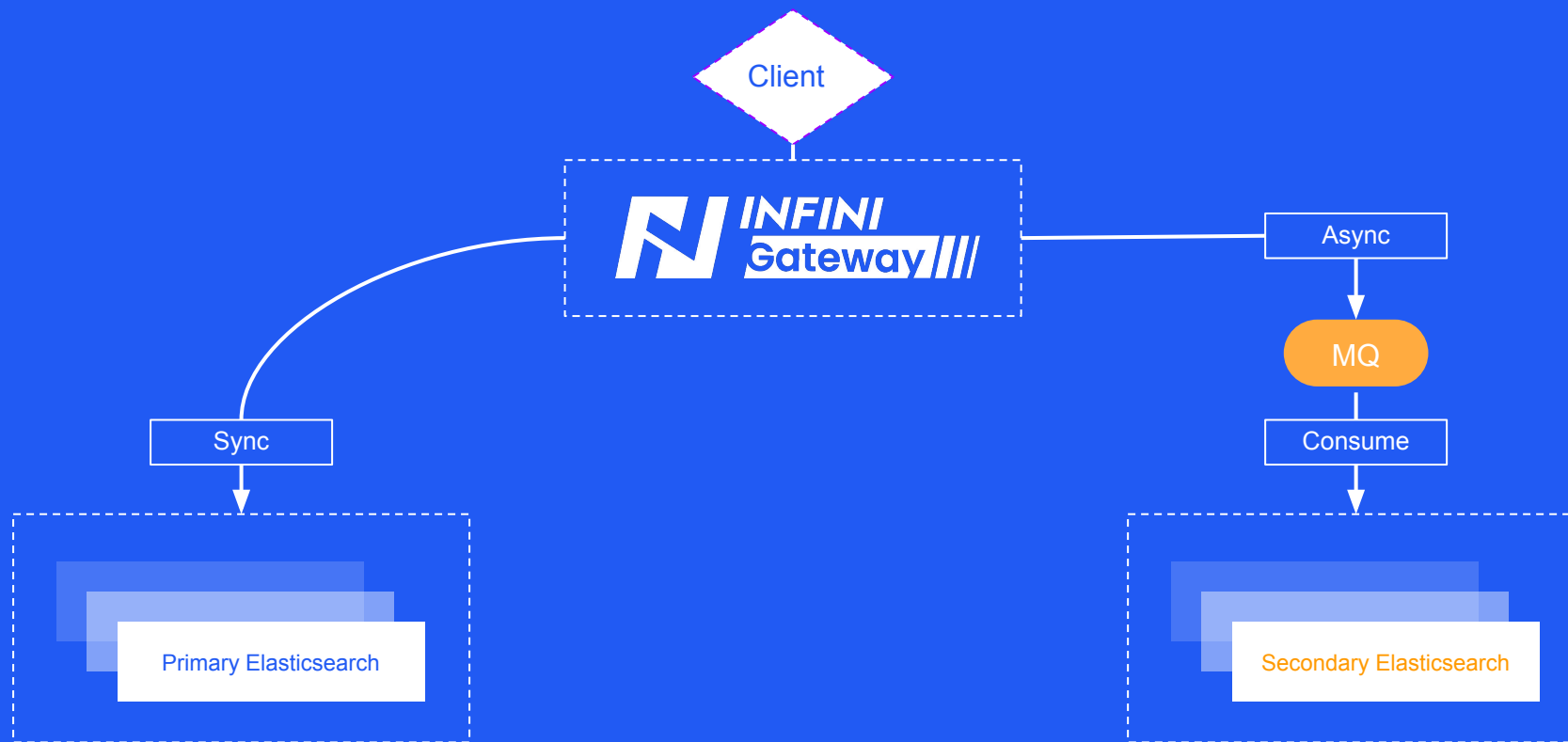
INFINI Gateway - Elasticsearch 专属网关



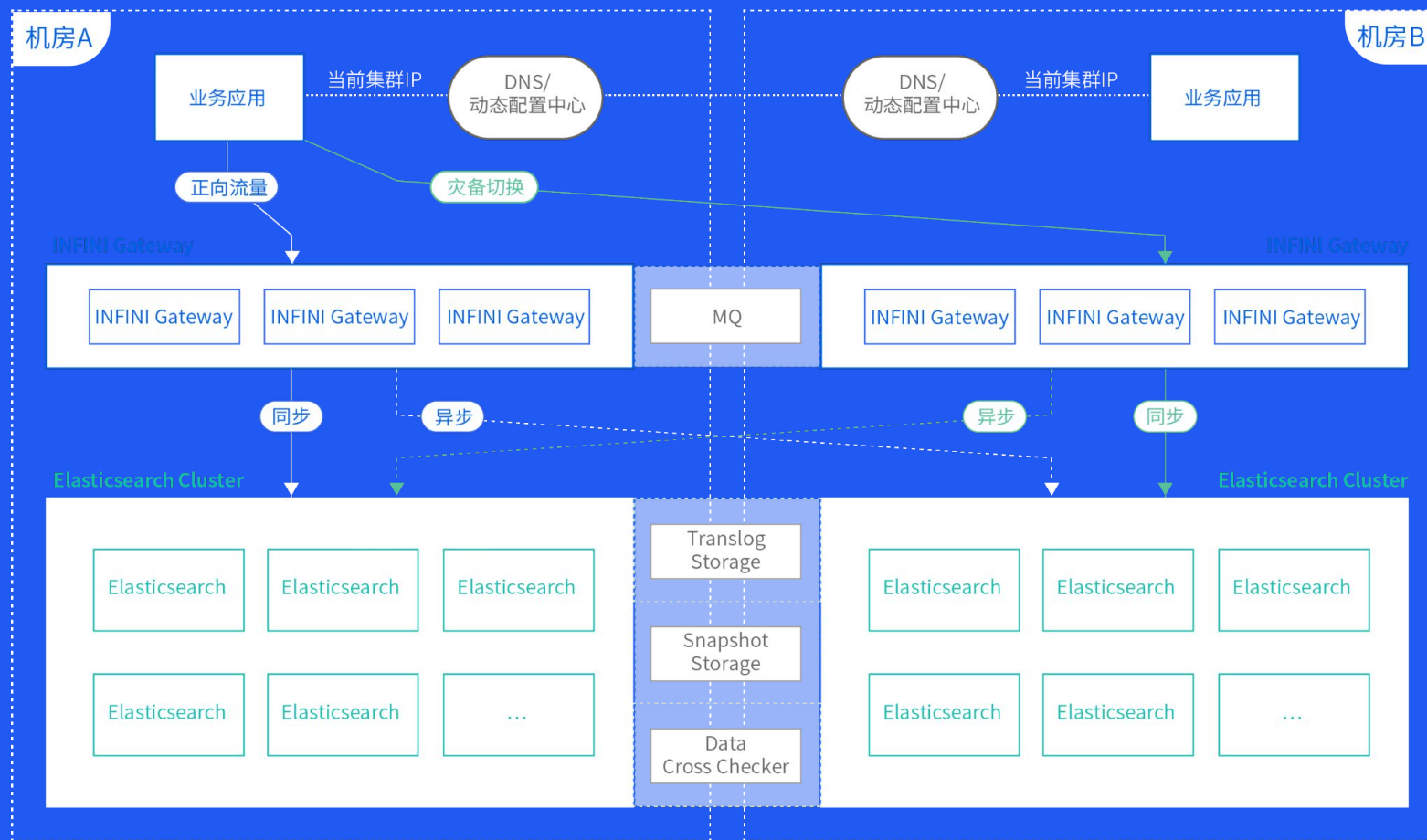


INFINI Gateway 容灾方案

INFINI Gateway 容灾架构



INFINI Gateway 容灾架构



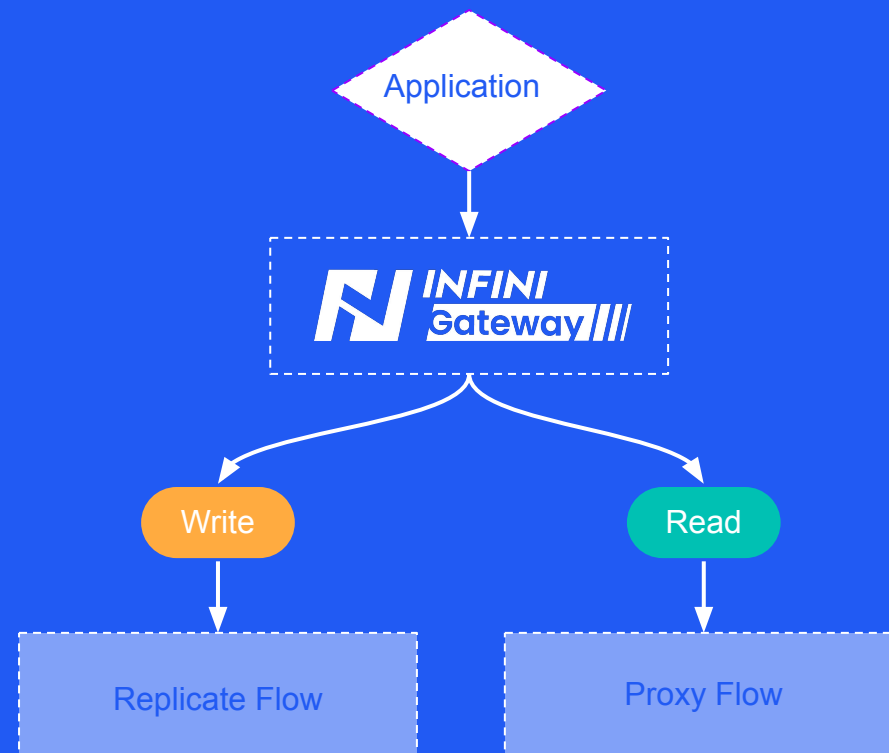
基于极限网关的 Elasticsearch 容灾设计

如何实现	服务的高可用	操作的顺序性
	数据的一致性	复制的时效性
副本的可验证	快速的可恢复	应用的灵活性



如何复制

- 基于文档操作进行复制
- 文档操作 API 稳定
- 可以跨版本进行复制
- 提前路由, 读写分离

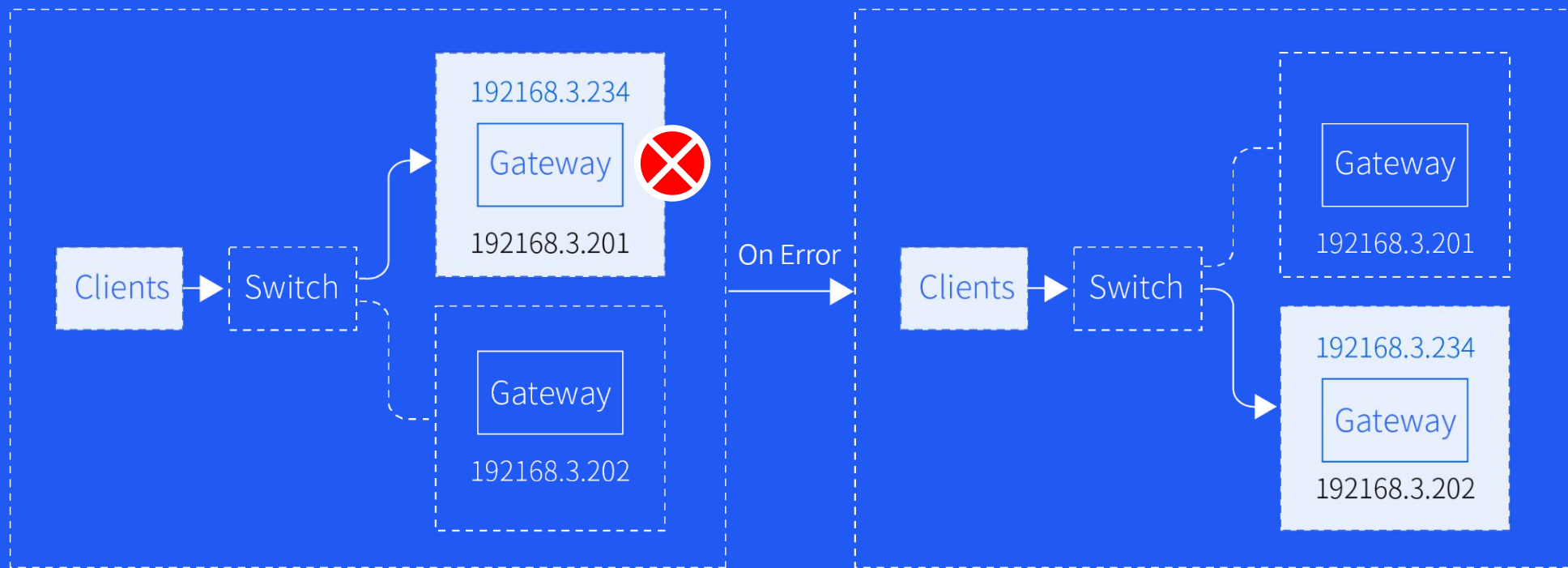




高可用如何保障？

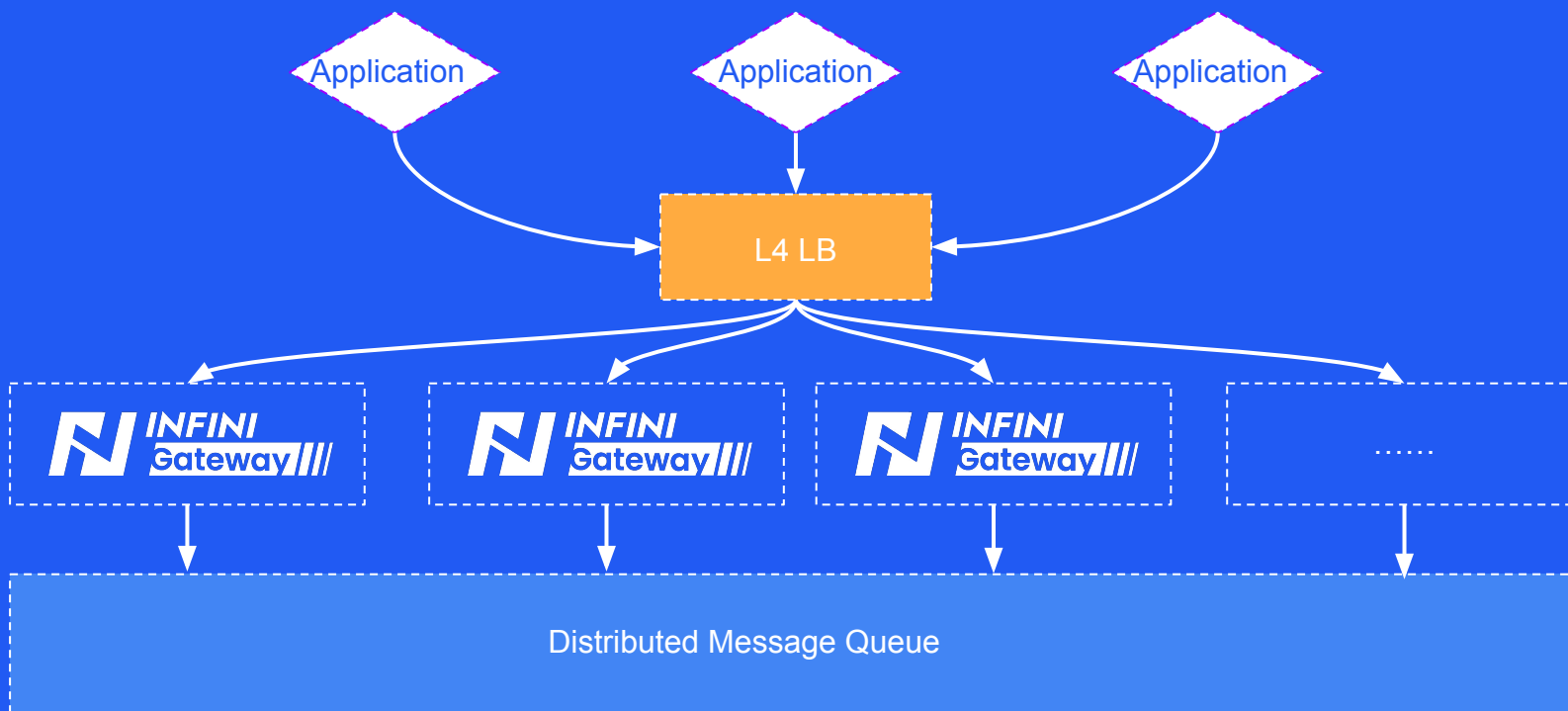
轻量级 - 双机热备模式

自带基于 VRRP 增强协议的虚拟浮动 IP 实现, 无需依赖额外组件



分布式-无状态大规模部署

动态水平扩容, 前置分拆流量, 存储依赖分布式消息队列





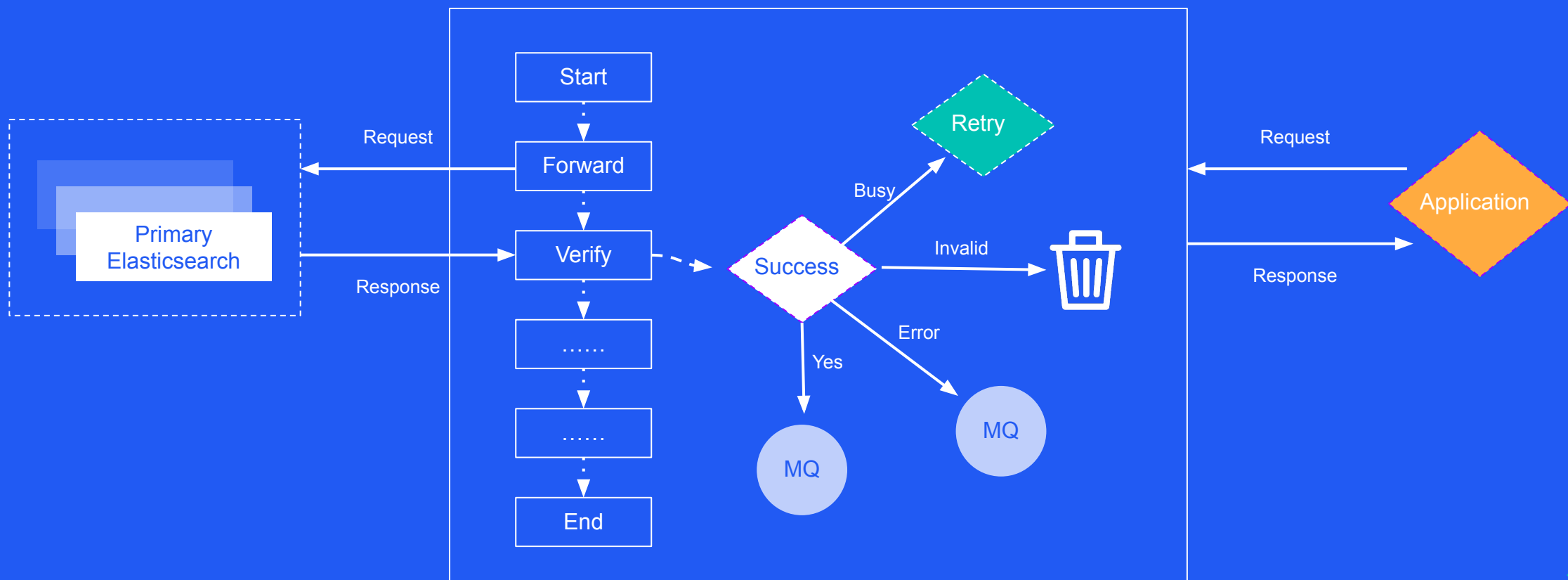
一致性如何保障？

双网关节点互备

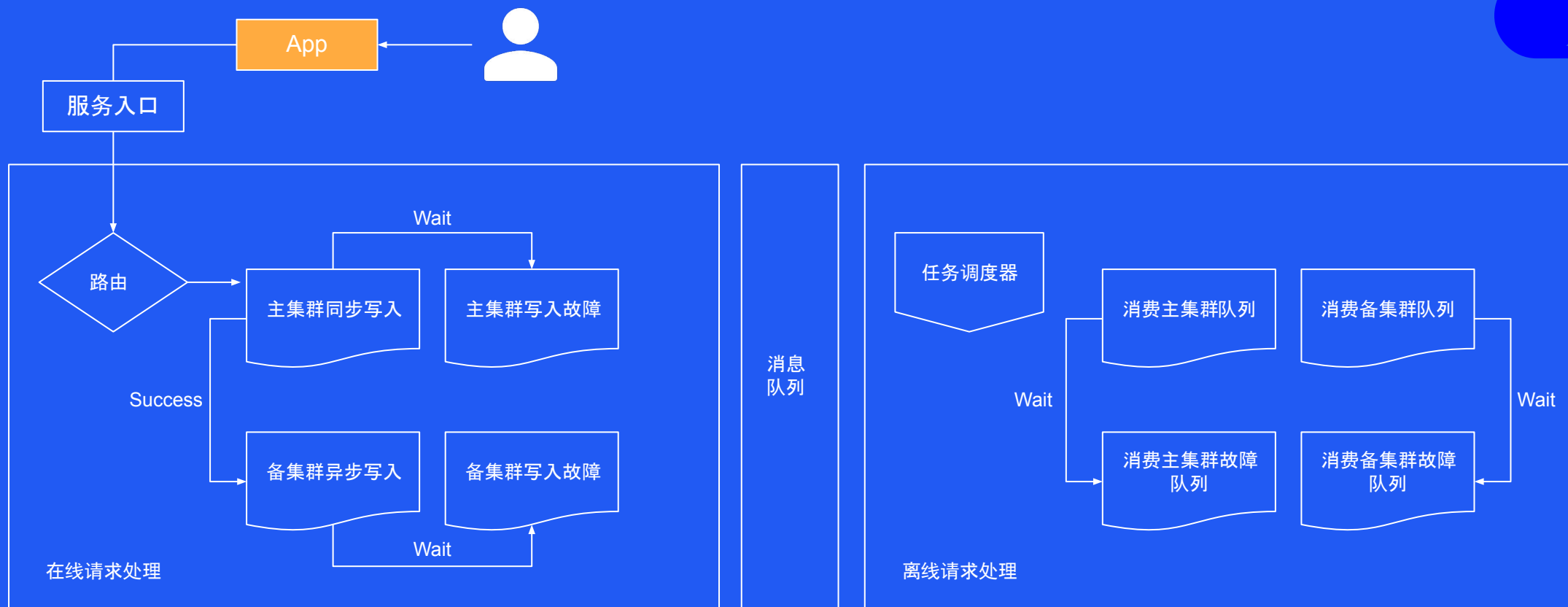
轻量双网关节点互备模式，本地磁盘队列时间线逻辑一致



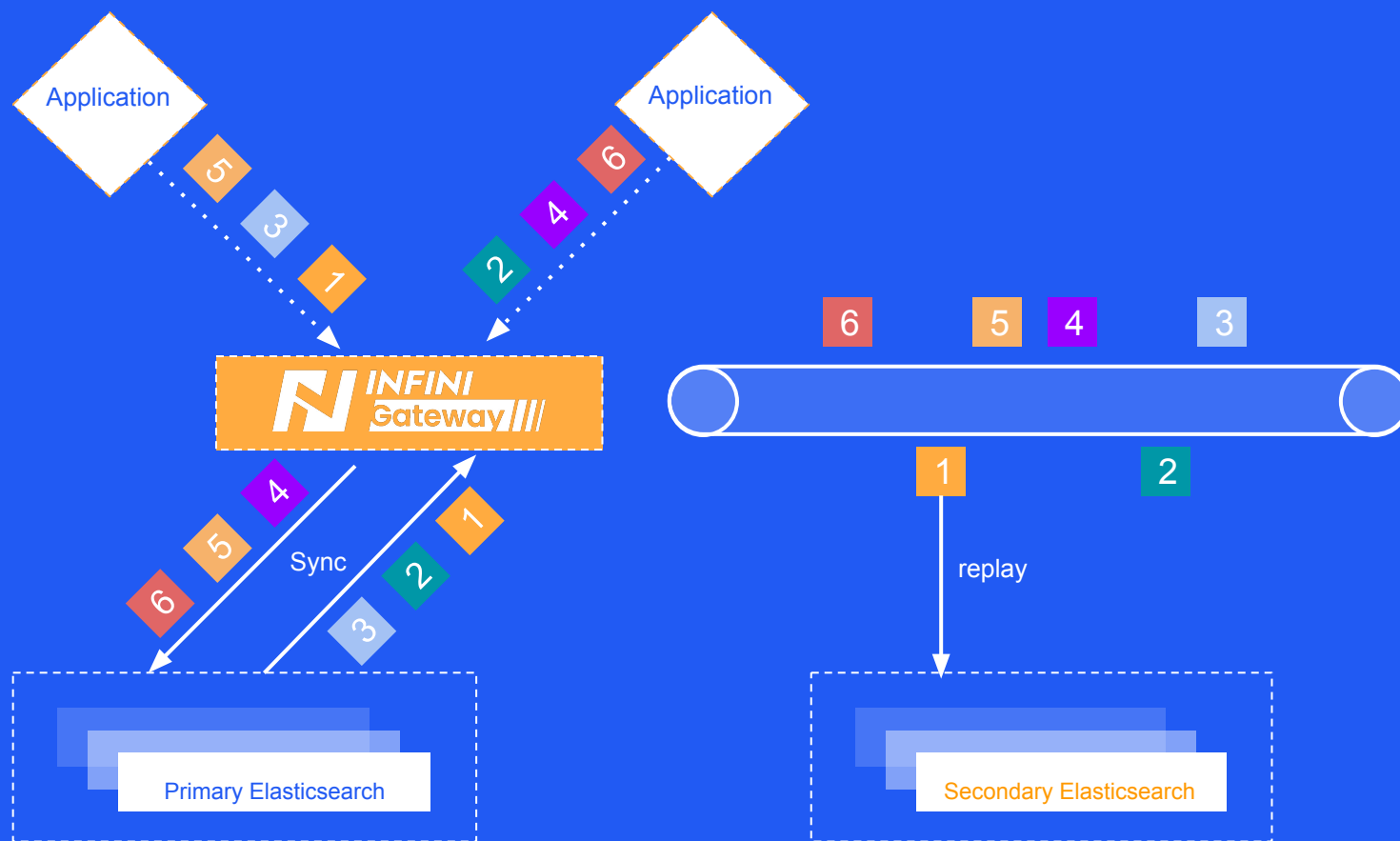
同步操作 校验返回



同步操作 校验返回



顺序入队 顺序重放

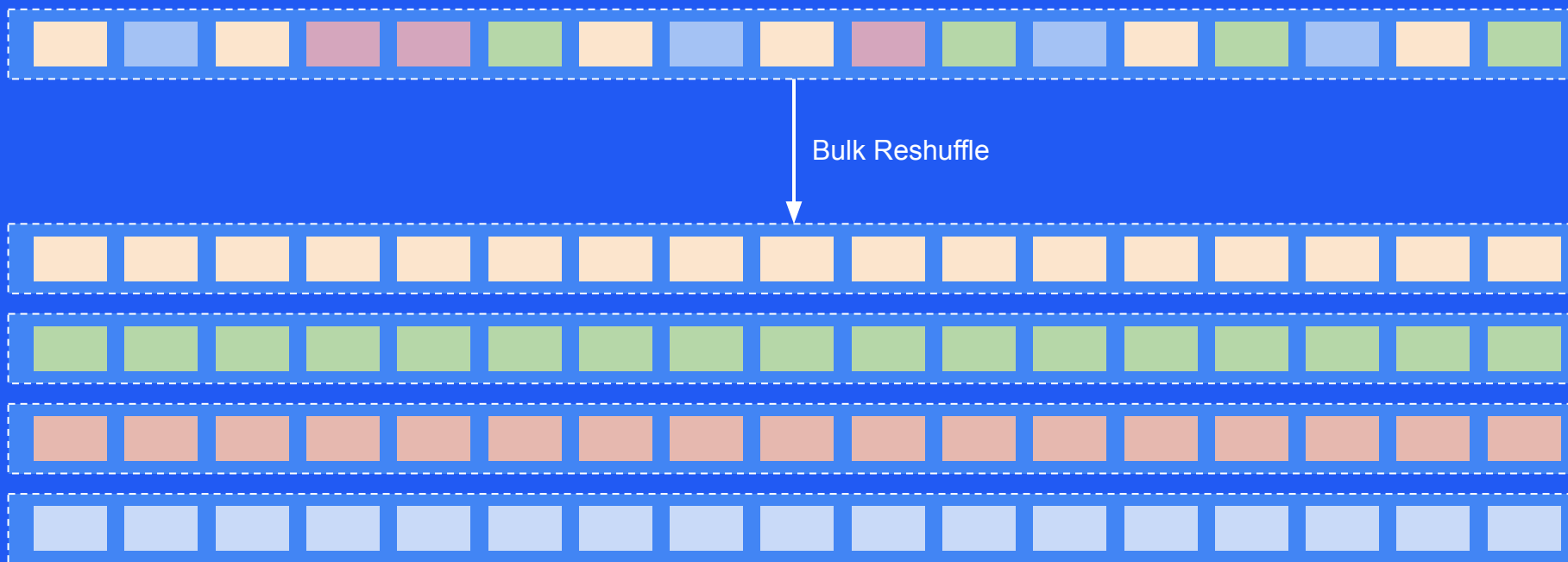




时效性如何保障？

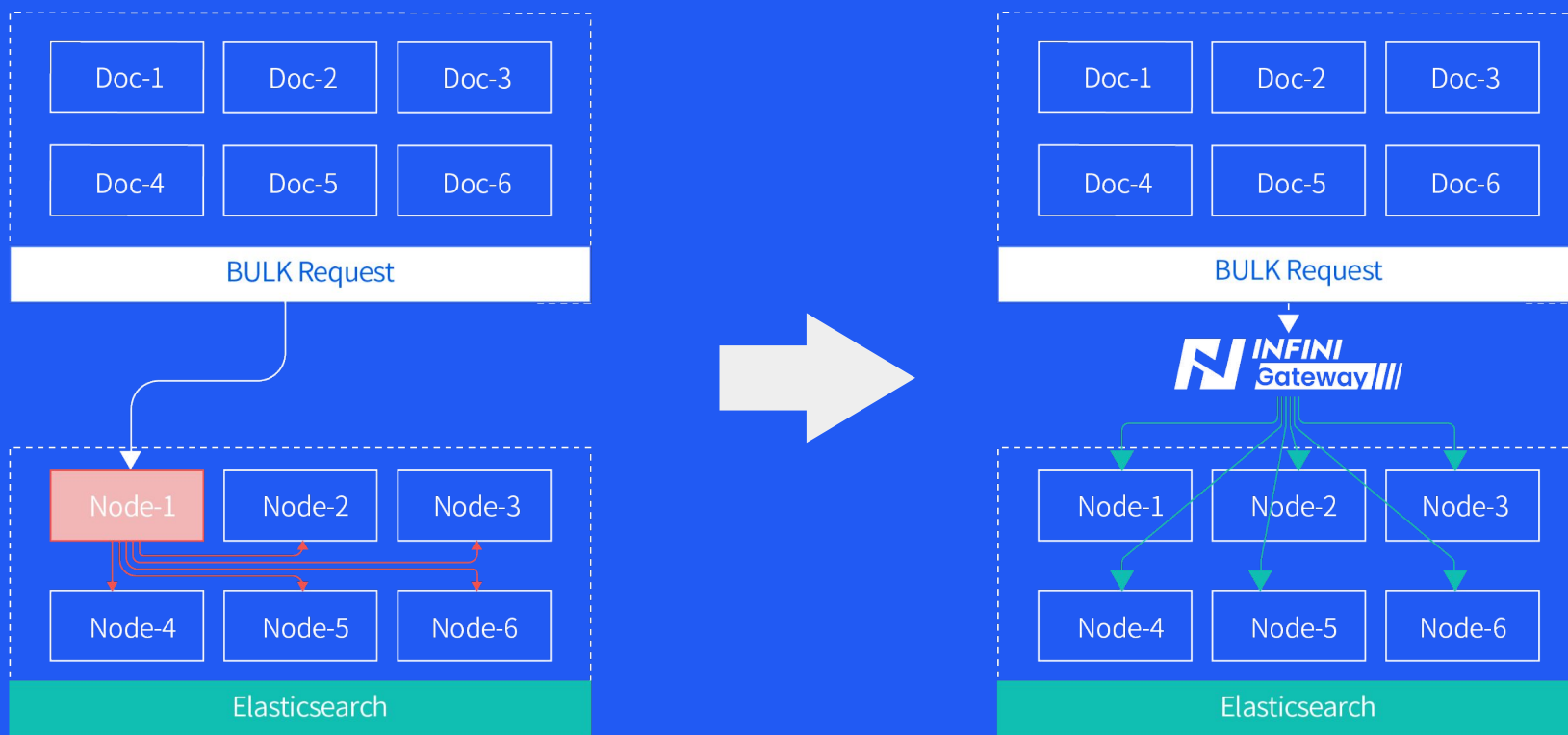
分拆合并

极限网关本地实现了 Elasticsearch 的非标 Hash 算法



定向投递

快慢分离, 稳定吞吐, 无缝提升 Elasticsearch 总体吞吐 30%~50%



一写多读

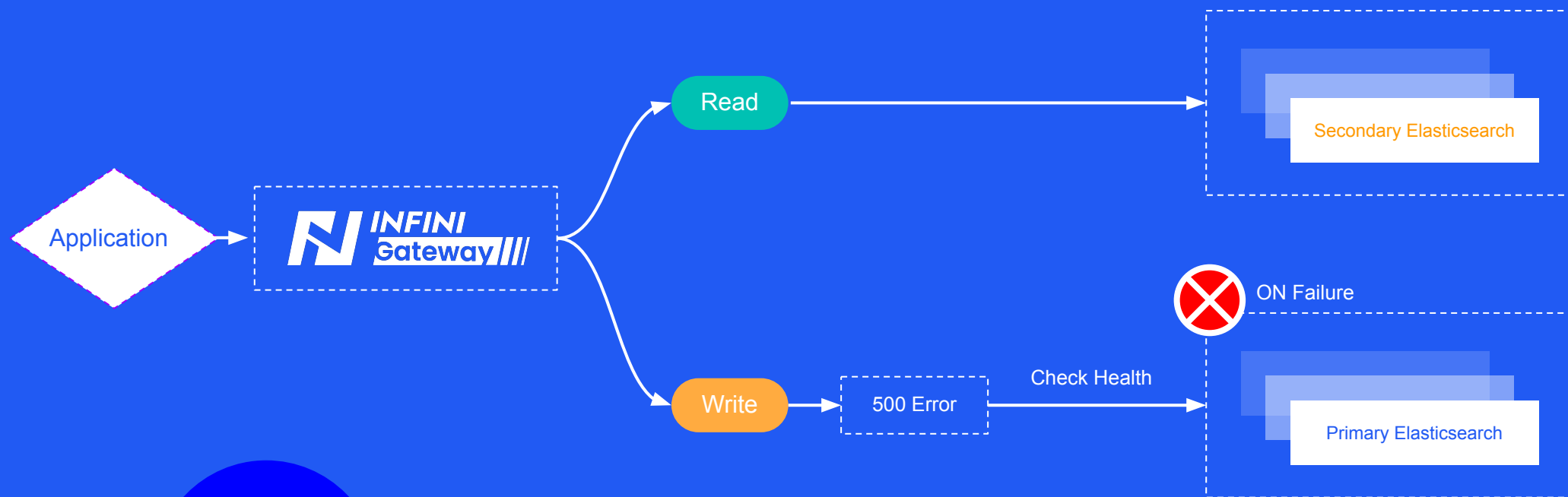
Runtime slice / Per slice consume



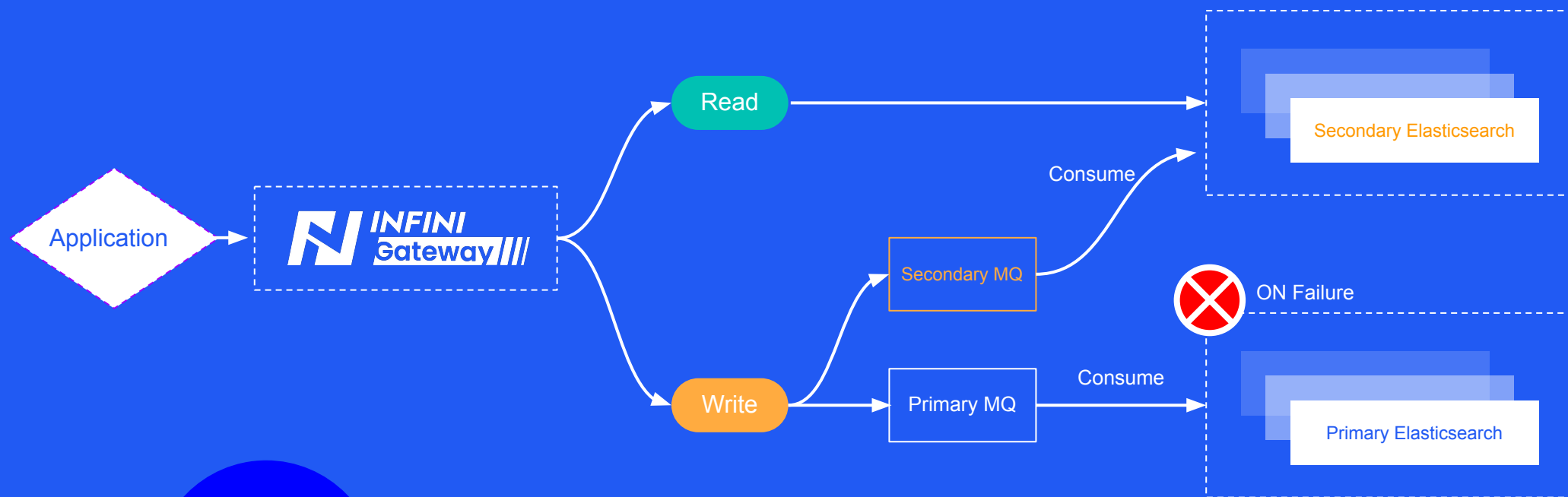
A large blue circle is positioned on the left side of the image. A dotted line, composed of small green dots, starts from the top right and extends diagonally towards the bottom left, passing through the circle. The text '容错能力呢?' is written in white, bold characters across the middle of the circle.

容错能力呢？

写入降级 查询不影响



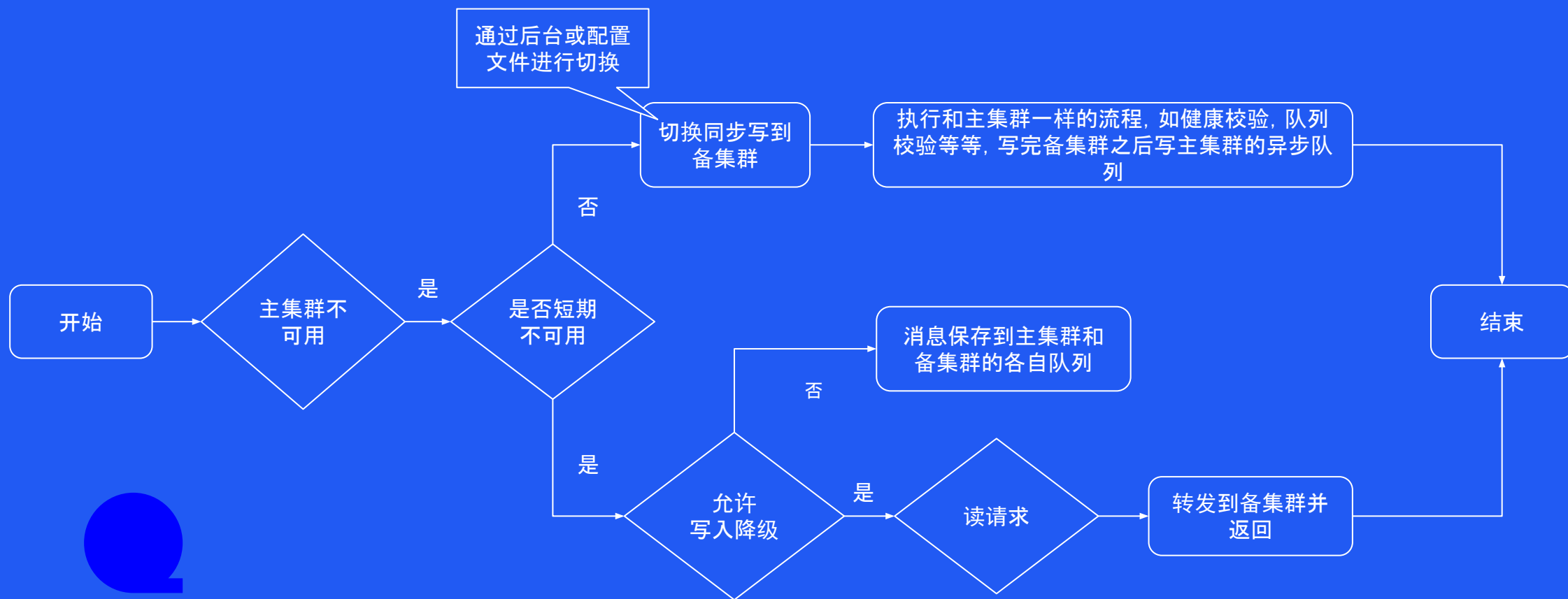
写入不中断 查询不影响



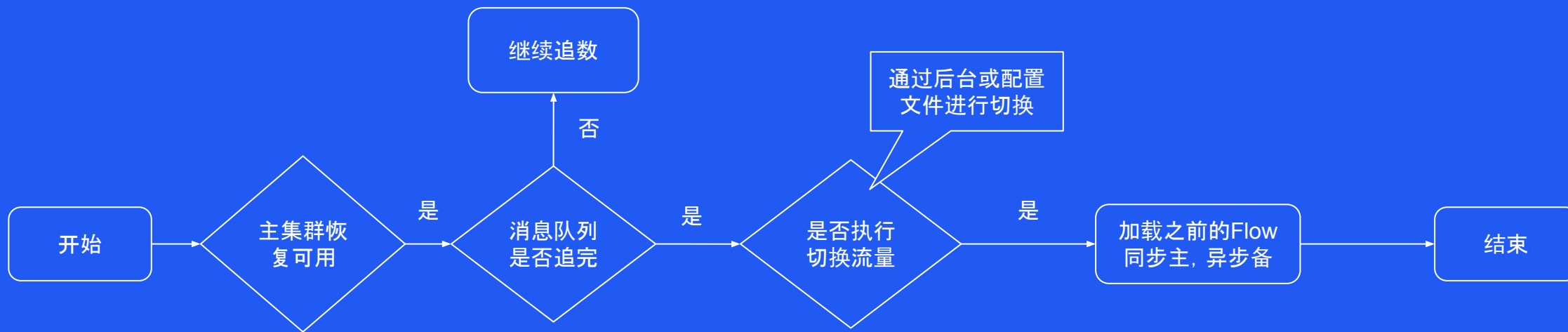


如何切换呢？

灾难主备切换流程



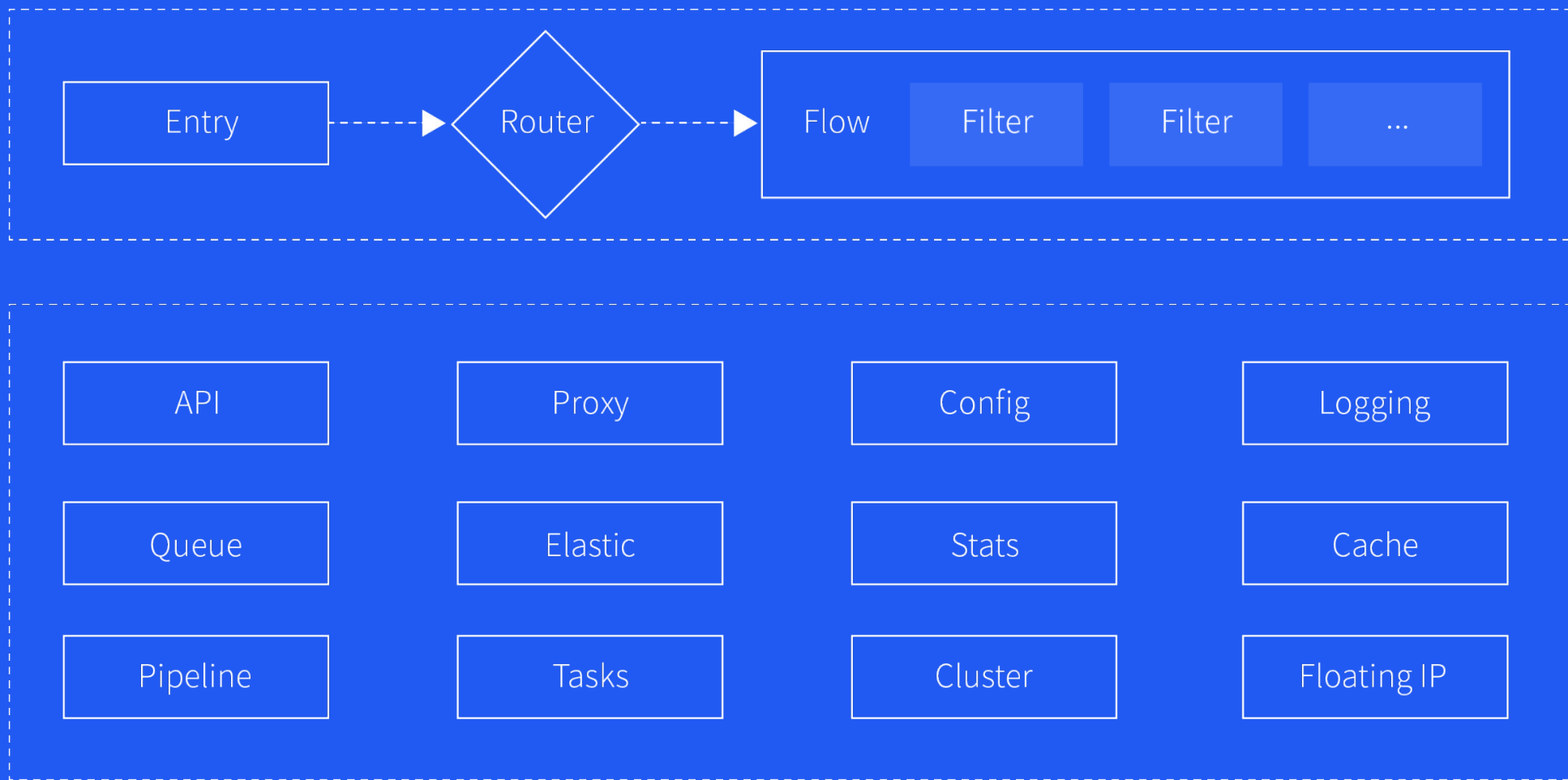
灾难恢复切回流程



A large blue circle is positioned on the left side of the image. A dotted line starts from the top right and extends diagonally towards the bottom left, passing through the circle. The text '灵活性怎么样?' is written in white, bold characters across the middle of the circle.

灵活性怎么样？

INFINI Gateway 核心模块



网关流程编辑

PLATFORM

ALERTING

DATA

GATEWAY

INSTANCE

ENTRY

ROUTER

FLOW

SYSTEM

Home / GATEWAY / FLOW / edit / UPDATE FLOW

编辑流程

修改流程配置(双击流程名可以修改流程名称)，然后点击保存按钮，保存成功之后生效。

primary-read-flow

GraphYaml

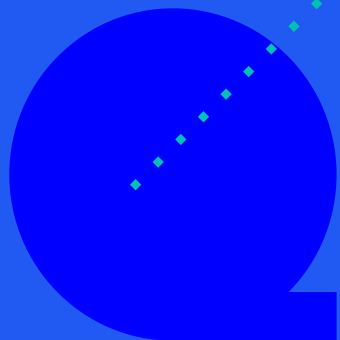
Save

INFINI Gateway 的服务处理单元 - Filter

<div>context_filter</div> <div>request_method_filter</div> <div>request_header_filter</div> <div>request_path_filter</div> <div>request_user_filter</div> <div>request_host_filter</div> <div>request_client_ip_filter</div> <div>request_api_key_filter</div> <div>response_status_filter</div> <div>response_header_filter</div> <div>echo</div> <div>dump</div> <div>record</div>	<div>ratio</div> <div>clone</div> <div>switch</div> <div>flow</div> <div>logging</div> <div>basic_auth</div> <div>ldap_auth</div> <div>queue</div> <div>elasticsearch</div> <div>cache</div> <div>translog</div> <div>redis_pubsub</div> <div>drop</div> <div>http</div>	<div>javascript</div> <div>sample</div> <div>request_body_json_del</div> <div>request_body_json_set</div> <div>context_regex_replace</div> <div>request_body_regex_replace</div> <div>response_body_regex_replace</div> <div>response_header_format</div> <div>set_context</div> <div>set_basic_auth</div> <div>set_hostname</div> <div>set_request_header</div> <div>set_request_query_args</div> <div>set_response_header</div> <div>set_response</div>	<div>context_limiter</div> <div>request_path_limiter</div> <div>request_host_limiter</div> <div>request_user_limiter</div> <div>request_api_key_limiter</div> <div>request_client_ip_limiter</div> <div>retry_limiter</div> <div>sleep</div> <div>date_range_precision_tuning</div> <div>bulk_resuffle</div> <div>elasticsearch_health_check</div> <div>bulk_response_process</div> <div>bulk_request_mutate</div>
--	--	---	--

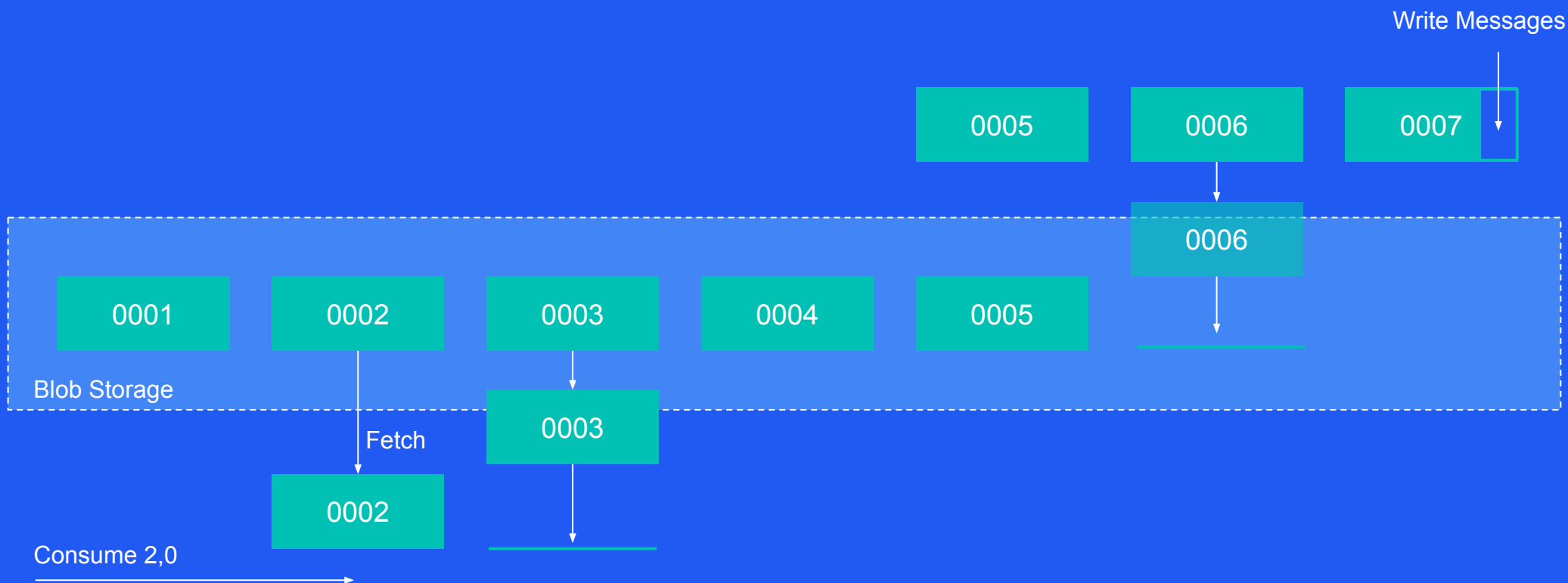


其他设计细节



INFINI Queue

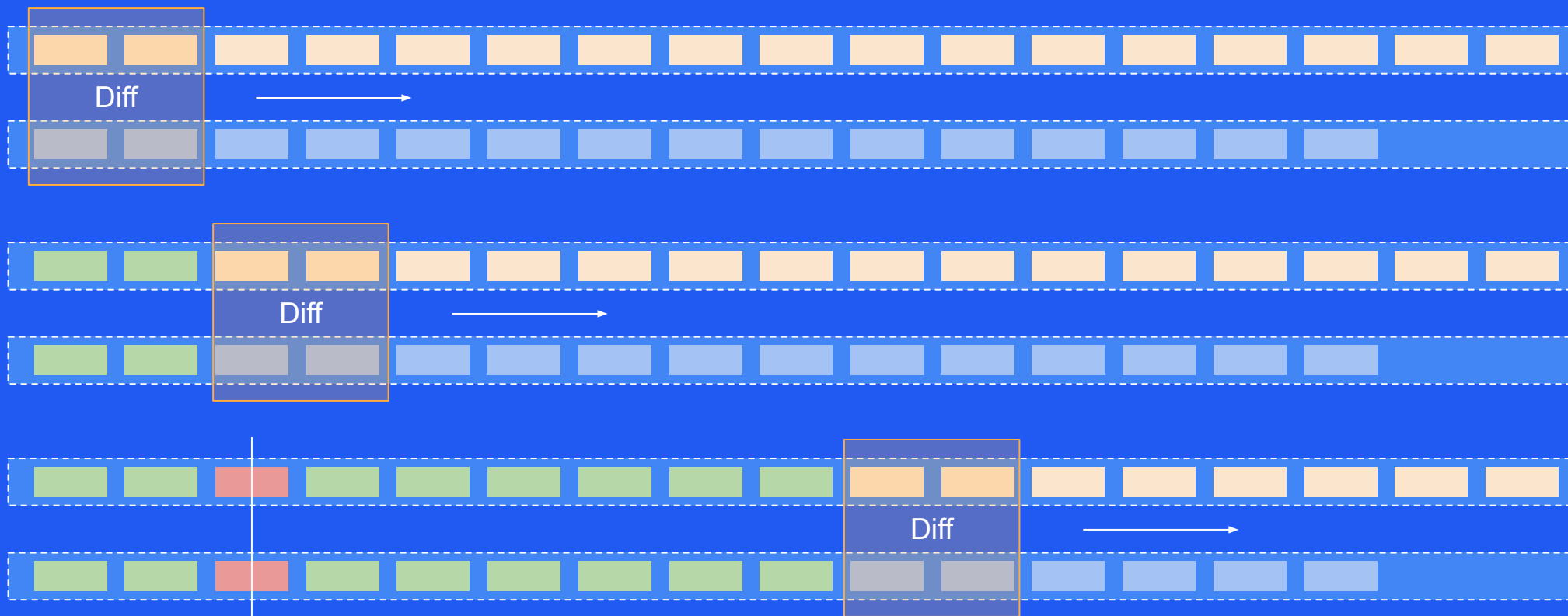
Cloud Native Lightweight Message Queue / 无限存储 存算分离



IndexDiff

实时增量

定期全量



Save Diff Result And Alerting!

数据迁移任务

PLATFORM

ALERTING

DATA

GATEWAY

SYSTEM

新建迁移任务

返回

迁移类型

☒ 全量+增量 ☐ 全量 ☐ 增量

迁移周期

15分钟

执行间隔

15分钟

运行时间

开始时间

至

结束时间

源集群

es-7140

Version:7.14.0

Indices:11

Docs:14.31M

Nodes:4

Shards:22

Disk Used: 999.01GB

目标集群

Es-710

Version:7.14

Indices:-11

Docs:1.21M

Nodes:-4

Shards:-1

Disk Used: 125.98MB

选择索引

选择迁移的索引

移除

您选择将要迁移的索引中，有2个索引在目标集群中已存在

☐ 手动处理 ☐ 覆盖 ☐ 不写入

您选择将要迁移的索引中，有2个索引在目标集群中已存在

☐ 手动处理 ☐ 覆盖 ☐ 不写入

选择迁移的索引

移除

数据迁移任务

[illegible]

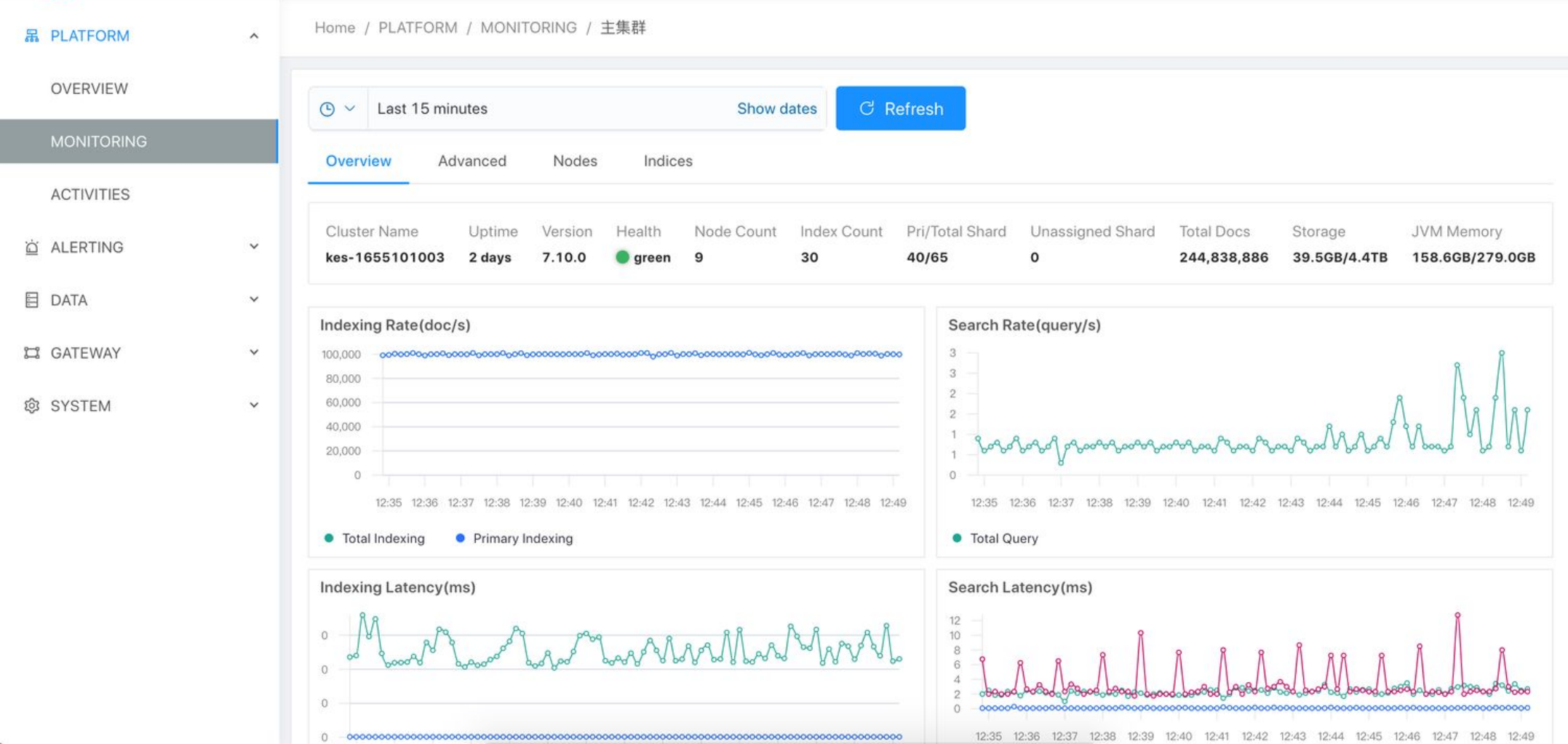


性能指标

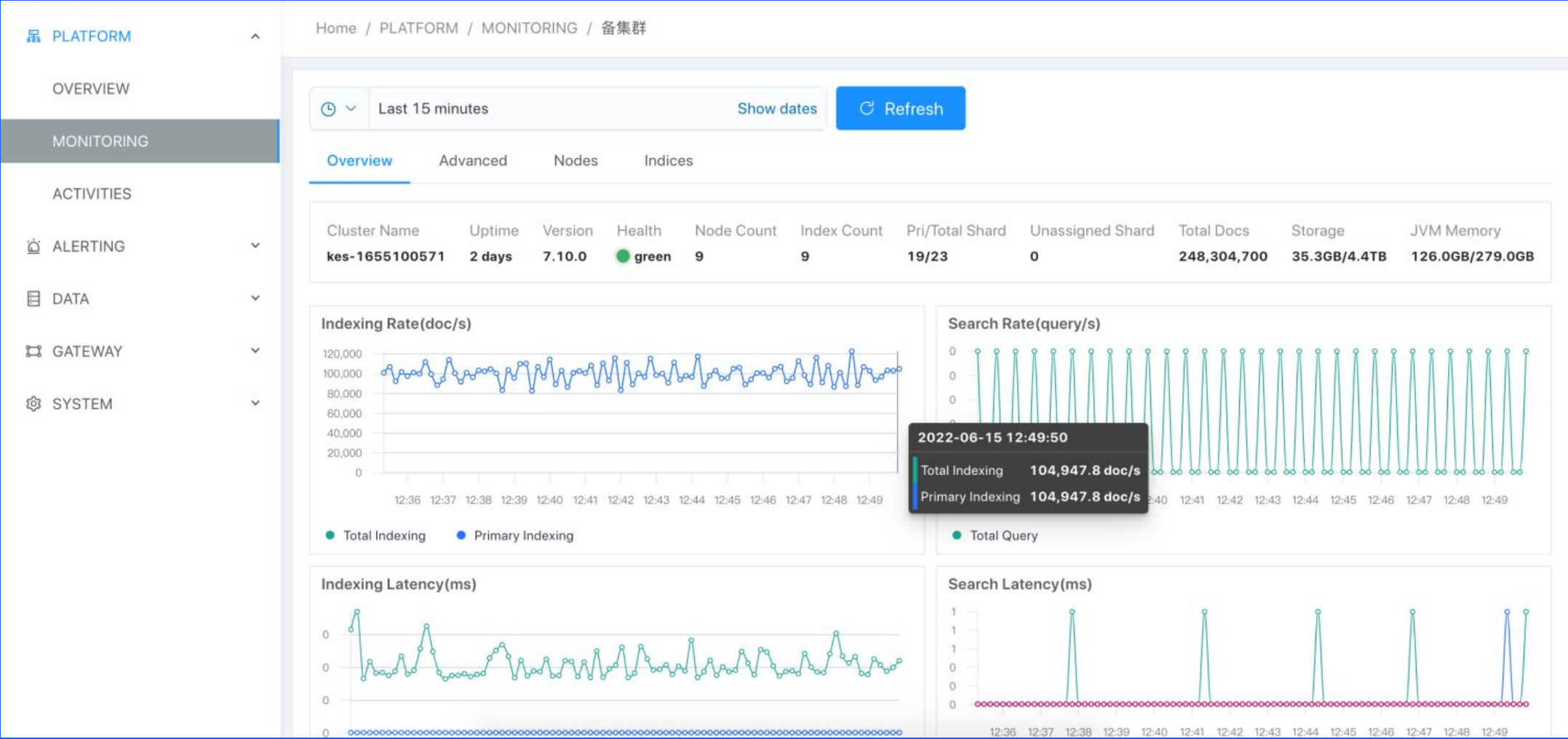
测试环境

- 主集群 : http: /10.0.1.2:9200, 用户名 :elastic 密码:**, 9 节点, 硬件规格:12C 64GB (31 GB JVM)
- 备集群 : http: //10.0.1.15:9200, 用户名 :elastic 密码:**, 9 节点, 硬件规格:12C 64GB (31 GB JVM)
- 网关服务器 1(公网 P:120.92.43.31, 内网 P:192.168.0.24) 硬件规格:40C 256GB 3.7 T NVME SSD
- 压测服务器 1(内网 P:10.0.0.117) 硬件规格:24C 48GB
- 压测服务器 2(内网 P:10.0.0.69) 硬件规格:24C 48GB

强一致性业务场景（网关 1C）



强一致性业务场景（网关 1C）



ELK日志场景

网关CPU核心数	复制能力（events per seond）	内存	备注
1C	~80k	~8GB	
2C	~160k	~8GB	
4C	~320k	~8GB	
8C	~600k	~8GB	
16C	~750k	~8GB	后端 ES 处理能力已接近饱和
32C	~750k	~8GB	后端 ES 处理能力已接近饱和

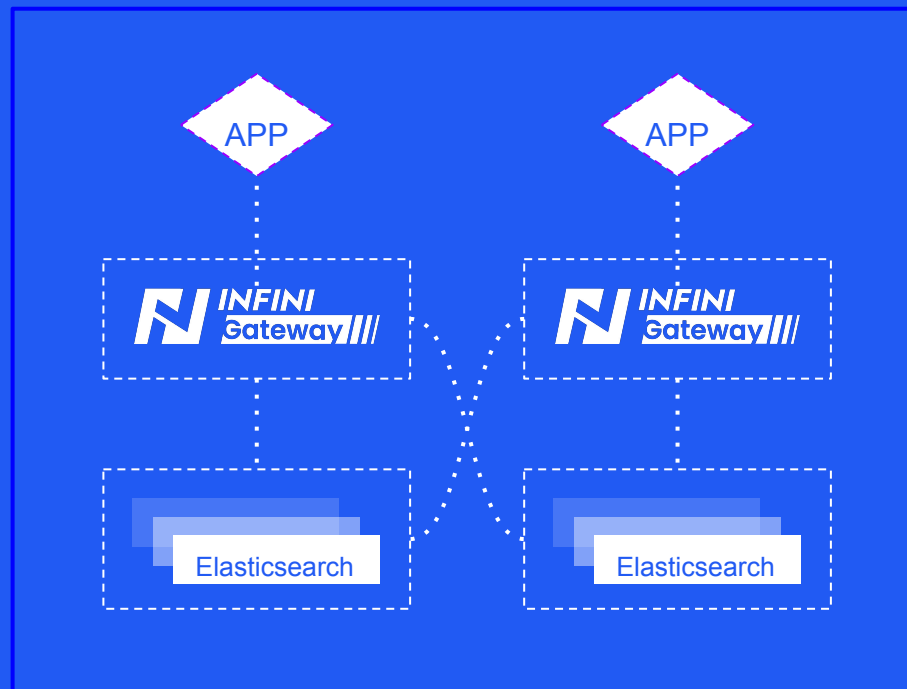
小结一下

功能

- 读写分离
- 请求级别 CDC
- 同步主
- 异步备
- 最终一致

优点

- 架构清晰简单
- 无缝透明, 应用无需任何调整
- 业务操作级别的复制, 跨版本兼容
- 双集群高可用, 随时切换
- 后端读写故障对前端业务无感知
- 节点故障自动处理, 请求不丢失
- 支持本地磁盘队列和 Kafka
- 结合快照和 Translog 可以重做索引
- 通过校验任务确保三方数据完全一致
- 自带四层网络虚拟 IP 高可用






INFINI Gateway

容灾方案其他应用场景

其他应用场景



无缝数据迁移	无缝版本升级	无缝上云
多云备份	无缝索引重建	读写分离



Thanks !



 400 139 9200



NINFINI Labs
简单 · 易用 · 极致 · 创新