Erlang在云数据库的实践和挑战

褚霸 2017年5月

云上数据链路的挑战

- 数据库服务极其严苛的可用性要求
- 云上业务多样性带来的困难
- 超高并发带来的问题
- 运维实体从几十到几十万带来的问题
- 资源倾斜，造成的服务不稳定
- 低成本与高效率的矛盾

数据链路主要产品

ApsaraDB Proxy
• DB高可用 • 安全审计 • 负载均衡 • 读写分离

HTAP DB
• 分库分表 • 分布式事务 • PB级存储 • 在线分析

云上数据链路的挑战

解法：

高可用与服务质量保证

高可用

场景：

1. 硬件故障
2. 操作系统故障
3. 内部逻辑bug造成的假死

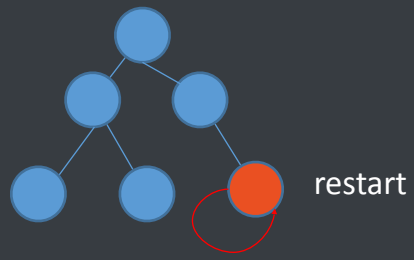
措施：

检查到异常后，将LVS路由摘掉，流量导走

高可用

故障时：

1. supervisor重启erlang process crash
2. 对于有状态的process，将状态保存到ETS，重启后恢复状态
3. heart 快速重启vm进程



高可用

小步快步,热代码替换：

1. 减少计划内的服务中断
2. 小步快跑，便于灰度
3. 逻辑变更与数据变更紧密配合
4. 相邻版本的兼容
5. 老代码监控

高可用

多种措施：

1. 降quota
2. Token buckets的流控算法
3. 安全审计变为抽查
4. 流量透传
5. 停止接收新连接

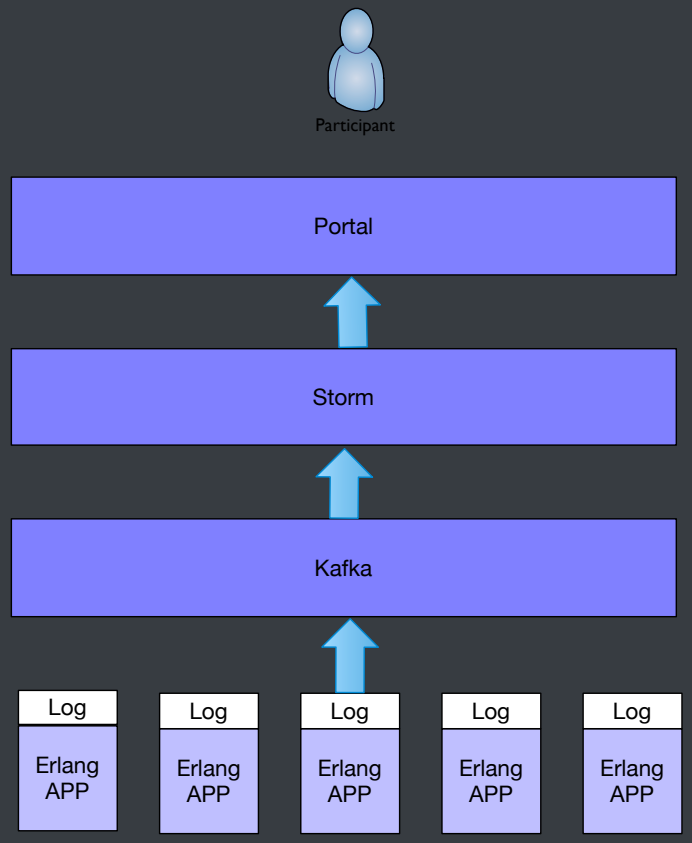
监控体系

采集与处理

数据源：
- recon
- 应用计数器
- system_info/1,memory/1
- etop

传输通道：
- 落地到带缓存的日志
- Logagent收集到kafka

处理分析：
- storm进行预处理
- PetaDB做汇总分析



指标维度

大类	子类	指标
系统指标	**调度器**	进程数、runq、util等
	内存	Gc次数，gc量，各分配器容
	网络	延迟，重传数，吞吐等
业务指标		连接数，qps，rt，各类异常
概况与日志		虚拟机配置信息，硬件和操作系统信息

性能优化

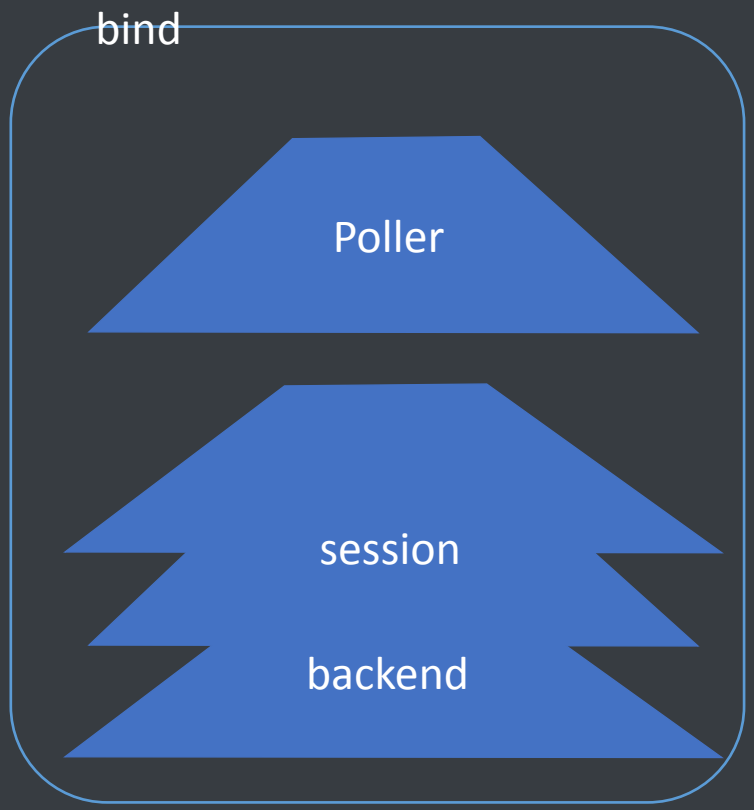
Profiling工具

- etop
- fprof/eprof/cprof
- systemtap
- perf
- warden
- 扁鹊

工欲善其事必先利其器!

小消息大处理

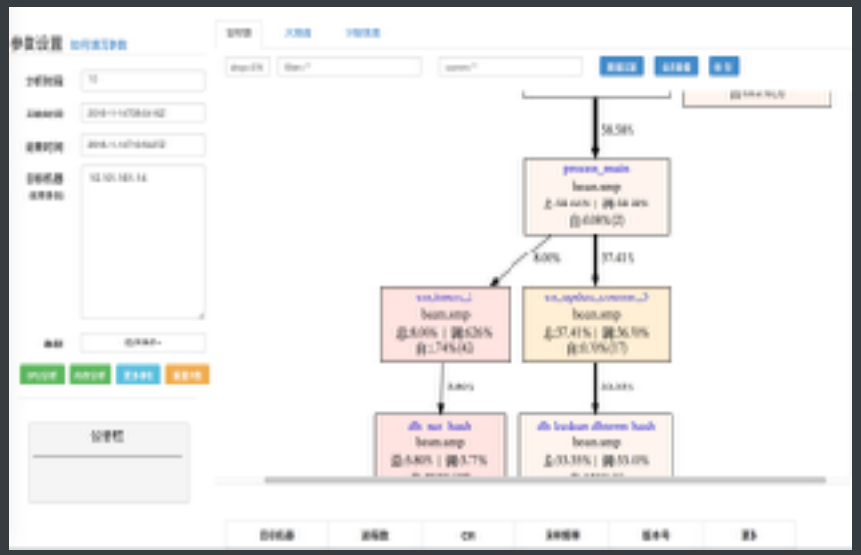
1. 避免大量的消息在不同的进程间传递——合并进程
2. 同一消息在同一scheduler——bind



扁鹊

功能：

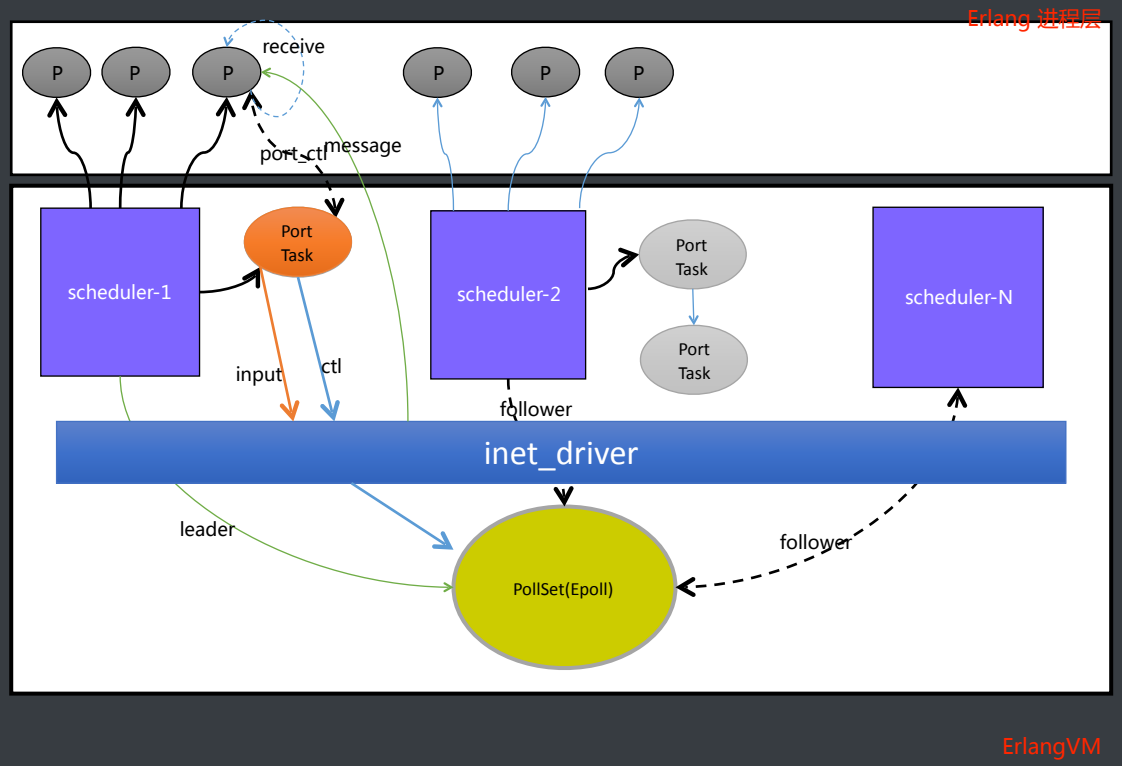
1. 收集内核态和用户态stack，函数调用频率信息
2. patch erts收集erlang层面调用链
3. 开销极小



Multi pollset

gen_tcp问题：

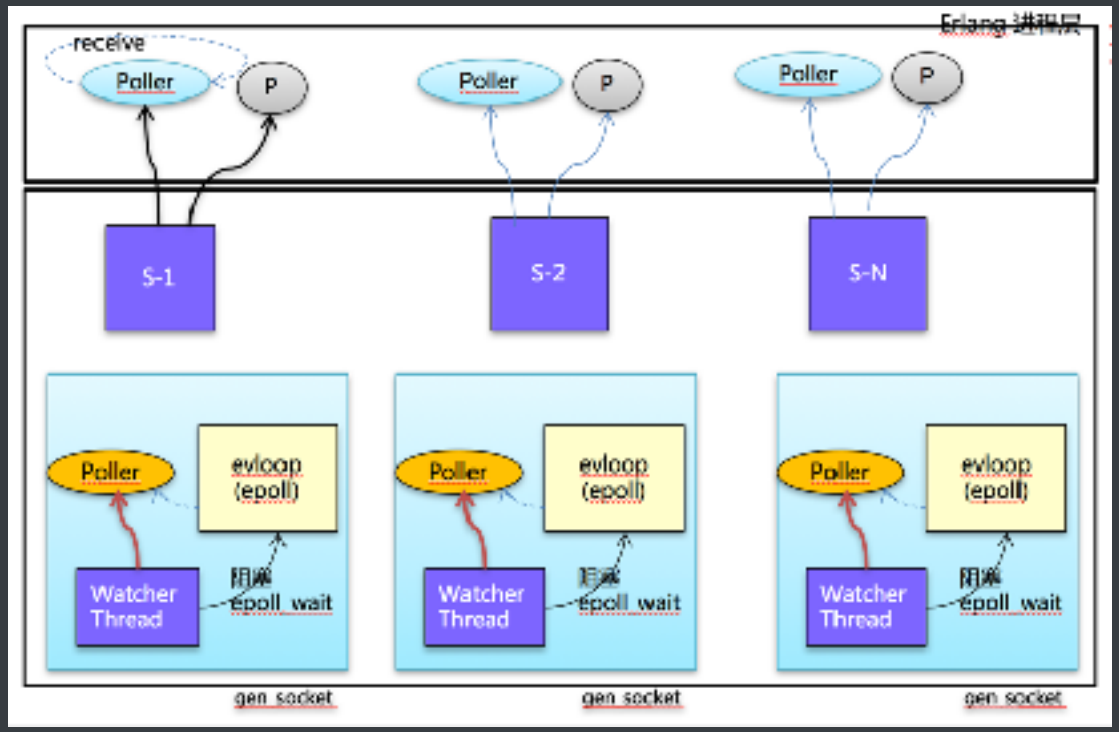
1. 单epoll set带来scale问题



Multi pollset

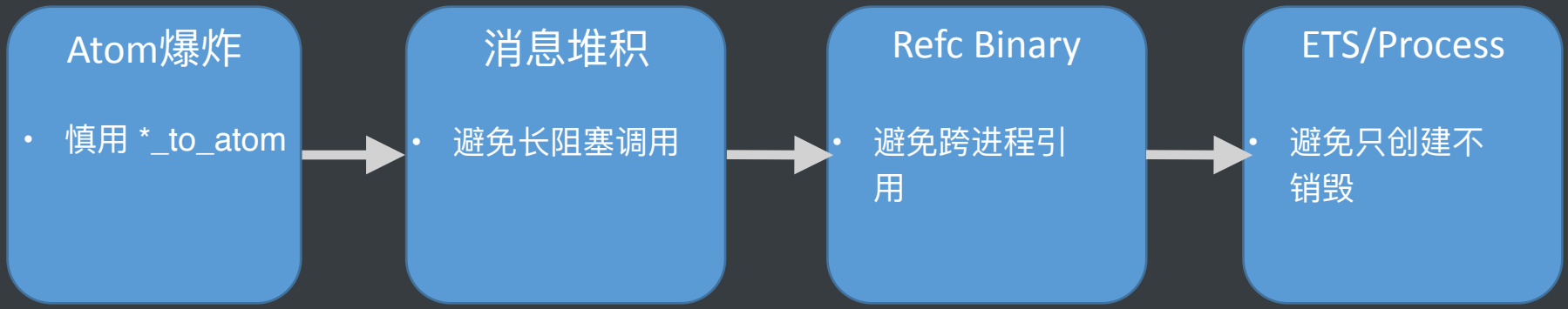
gen_socket:

1. N个OS线程收割网络事件
2. 对调度器,用户进程透明
3. 接口和gen_tcp完全兼容。
4. 细粒度锁优化
5. 性能提升110%
6. 已开源： https://github.com/alibaba/erlang_multi_pollset



问题与解决之道

内存泄漏


Atom爆炸
• 慎用 *_to_atom
消息堆积
• 避免长阻塞调用
Refc Binary
• 避免跨进程引用
ETS/Process
• 避免只创建不销毁

调度不均衡

NIF引起的	大Io引起的
拆解nif函数为小操作，定期调用 **enif_consume_timeslice** **add reductions**	启用dirty scheduler，配置适当的数量

热升级注意事项

1. 尽量soft_purge，purge不成功不要强行替换
2. 数据与应用逻辑的兼容，写好code_change
3. 避免模块间循环依赖

欢迎加入

阿里云ApsaraDB团队的工作范围涉及到OS内核(资源隔离)、存储、引擎（TP、AP）、数据库内核、中间件、管控、监控、数据流动、计算服务等，是个复杂和精美的协作团队, 已经有10几个数据库相关产品，在市场地位和收入上都有不错的表现，团队有业内非常有经验的人，欢迎大家加入，团队主力主要在杭州和北京