Software Engineering Conference Russia

14-15 ноября, 2019. Санкт-Петербург

Автоматический выбор оптимального набора журналов для отчетов об ошибках

Денис Силаков

Virtuozzo

ПО без ошибок — подозрительное ПО

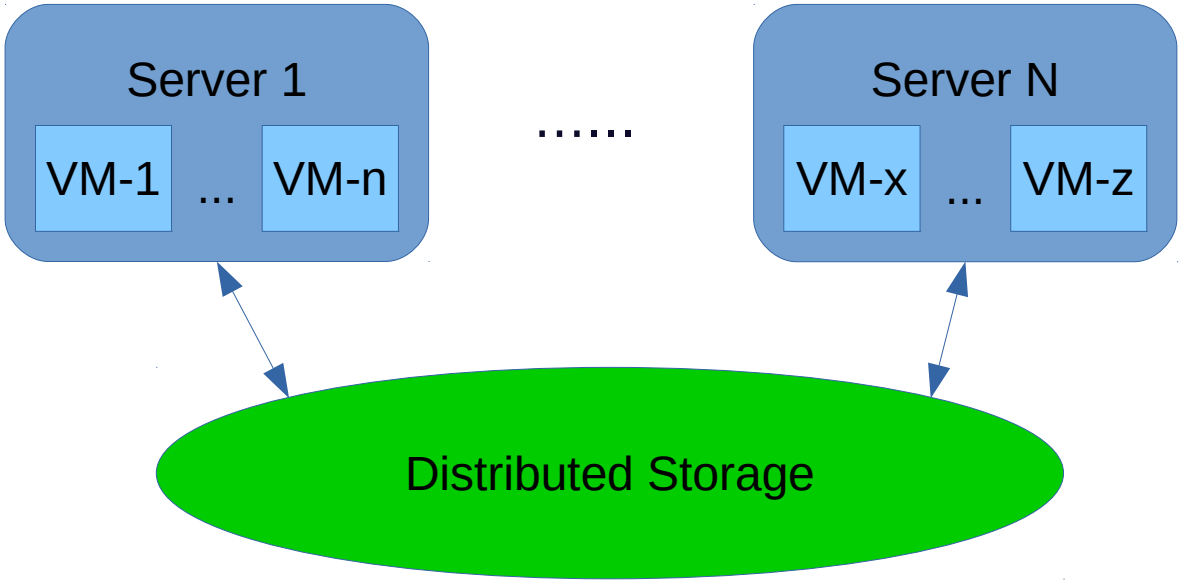
Понимание причины проблемы — половина исправления

Отчет об ошибке:

1. Снимок памяти процесса
2. Stack Trace (Call Trace) - последовательность вызовов функций, приведшая к проблеме
3. Журналы
4. ...

Как выбрать журналы?

Virtuozzo:



1. Все подряд?
   - Слишком много *(~400 Gb в день)*
2. В зависимости от упавшего процесса
   - Слишком грубо *(~300 Gb в день)*
3. <u>В зависимости от трассы</u>

Кластеризация ошибок

1. Набор эталонных трасс
   - С известным набором журналов для каждой
2. Оценка схожести трассы с эталонным набором
   - Выбор тех, с кем близость выше порога

Оценка схожести

(1) **f1** → f2 → f3

(2) f2 → **f1** → f3

(3) **f1** → f2 → f3 → f4 → f5 → f6

1. Общий префикс
   - (1) vs (2) — ничего общего
2. Сравнение строк (String Edit Distance)
   - (1) vs (2) — 1 перестановка, (1) vs (3) — 3 вставки
3. Гибридные методы
   - Все функции имеют значение
   - Кто ближе к началу трассы — тот весомее

Position Dependent Method

Из работы «ReBucket: A method for clustering duplicate crash reports based on call stack similarity»

D(f) — расстояние до начала трассы (минимум по трассам)

S(f) — разница позиции в трассах

L_1, … L_n — общие подпоследовательности функций

$$Q(L_i) = \sum_{f \in L_i} e^{-cD(f)} \times e^{-oS(f)}$$

$$sim(C_1, C_2) = \frac{max_{L_i \in L}(Q(L_i))}{\sum_{j=1}^{l} e^{-cj}}$$

Сложность: N^2

PDM в действии

(1) **f1** → f2 → f3

(2) f2 → **f1** → f3

(3) **f1** → f2 → f3 → f4 → f5 → f6

1. (1) vs (3) — одинаковы
2. (1) vs (2):

$$sim(C_1, C_2) = \frac{e^{-c} \cdot e^{-o} + e^{-3c}}{e^{-c} + e^{-2c}}$$

Обучение

Параметры:

o & **c** для PDM, **r** — порог для отбора трасс

1. Перебор сочетаний параметров
2. Вычисление **F1-score** для каждого
3. … и выбор сочетания с наибольшим F1-score

$$F_1 = \frac{2 \times TruePositive}{2 \times TruePositive + FalseNegative + FalsePositive}$$

Препроцессинг

1. Избавление от рекурсии
2. Добавление имени процесса в трассу

   f0 = process_name
3. Унификация схожих функций
4. Удаление «точно хороших» функций

Virtuozzo - апробация

Virtuozzo 7, Virtuozzo Infrastructure Paltform 3

1. Обучающая выборка: **2.000** трасс,

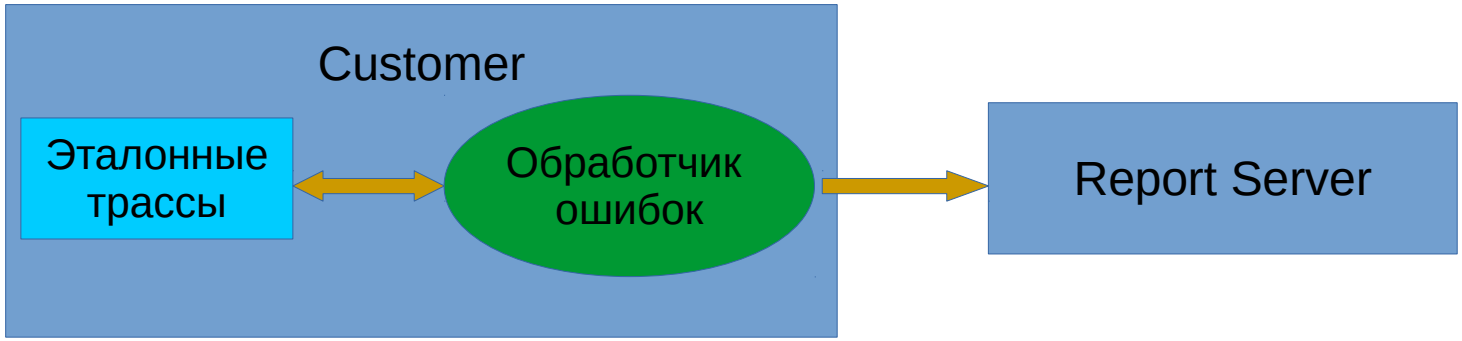
   ~50 потенциально нужных журналов

2. Тестовая выборка: **2.500** трасс
   - *Потеряно **2** нужных файла*
   - *Для **~100** трасс были добавлены лишние файлы*

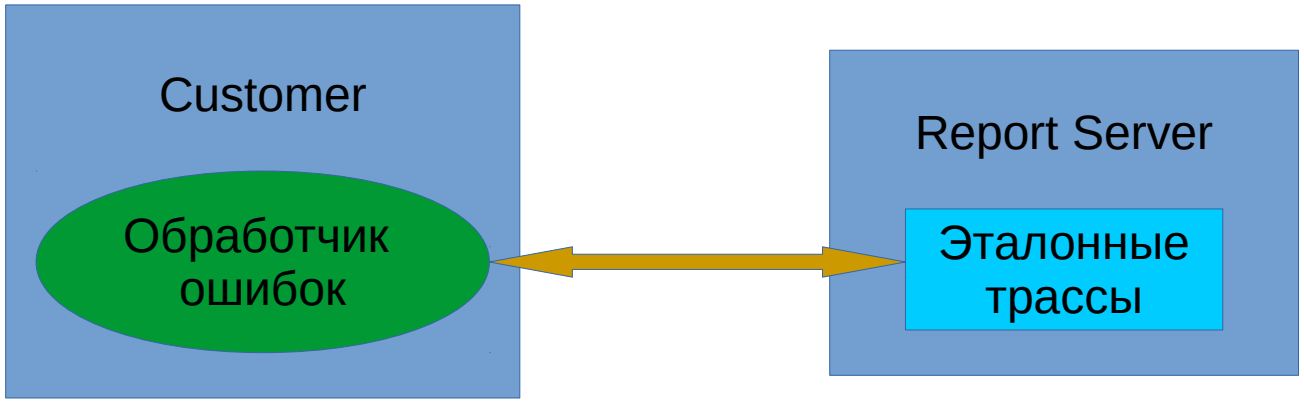
3. Выигрыш: **~10%** по сравнению с текущим подходом (**20-30Gb в день**)

Внедрение

1.



2.



Контакты

Денис Силаков

Virtuozzo

dsilakov@virtuozzo.com