Ranger

Streaming based tracking & monitoring system for cloud services

Agenda

- Palo Alto Networks
- Wildfire Cloud Service
- Pain points on DevOps
- Ranger—streaming based tracing & Monitoring system
  - Goal
  - Solution
- Future work
- Questions

Abstract

- This presentation will discuss the pain points for DevOps operation at Palo Alto Networks managing a cloud system and address how to utilize streaming technology to design and implement the real time tracing and monitoring system. How to use streaming pipeline to calculate cloud service performance matrix based on application business logic, use Elasticsearch to store matrix data, customized GUI or Grafana to present the statistic result and entity process tracing result. And to discuss the future work being done on this streaming system to make it serve as a smart to assistant for DevOps daily operation.

Palo Alto Networks

About Palo Alto Networks

- Palo Alto Networks is the next-generation security company, leading a new era in cybersecurity by safely enabling applications and preventing cyber breaches for tens of thousands of organizations worldwide. Built with an innovative approach and highly differentiated cyberthreat prevention capabilities, our game-changing security platform delivers security far superior to legacy or point products, safely enables daily business operations, and protects an organization's most valuable assets.

- http:// www.paloaltonetworks.com

Palo Alto Networks

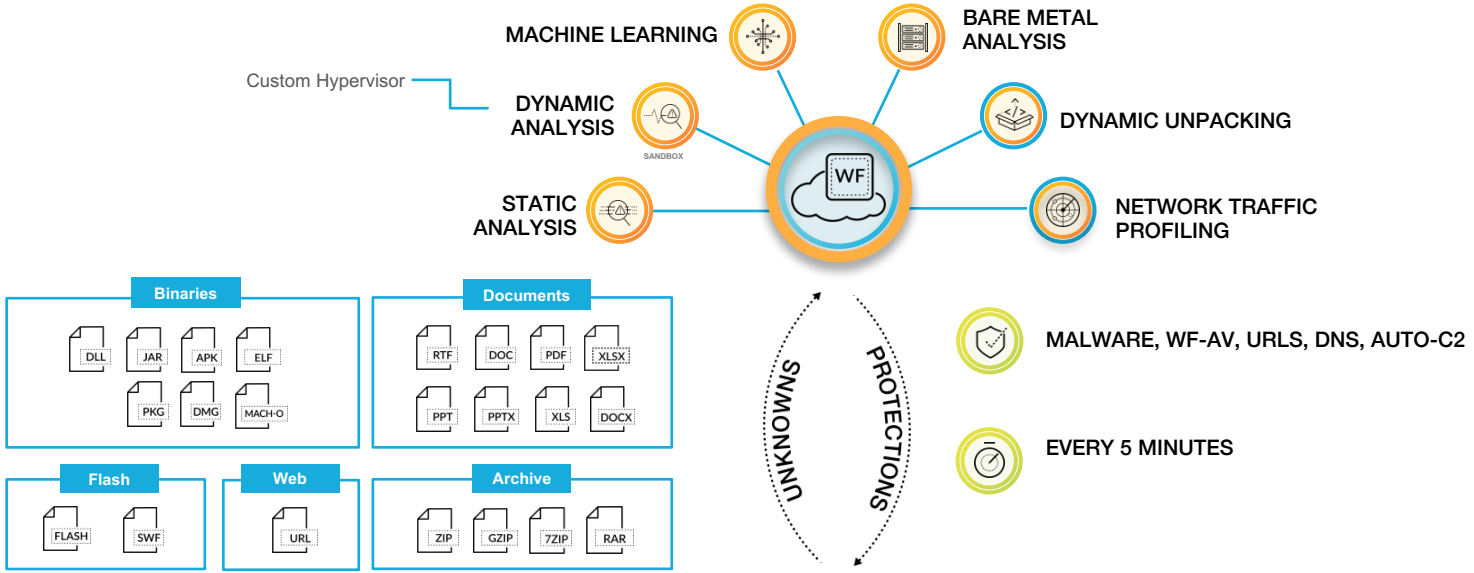
- Upholds Leadership position in Gartner Magic Quadrant for Enterprise Networks Firewalls

Wildfire Cloud Service

Wildfire Cloud Service @ PaloAlto networks

- *WildFire® cloud-based threat analysis service is the industry's most advanced analysis and prevention engine for highly evasive zero-day exploits and malware. The cloud-based service employs a unique multi-technique approach combining dynamic and static analysis, innovative machine learning techniques, and a groundbreaking bare metal analysis environment to detect and prevent even the most evasive threats.*

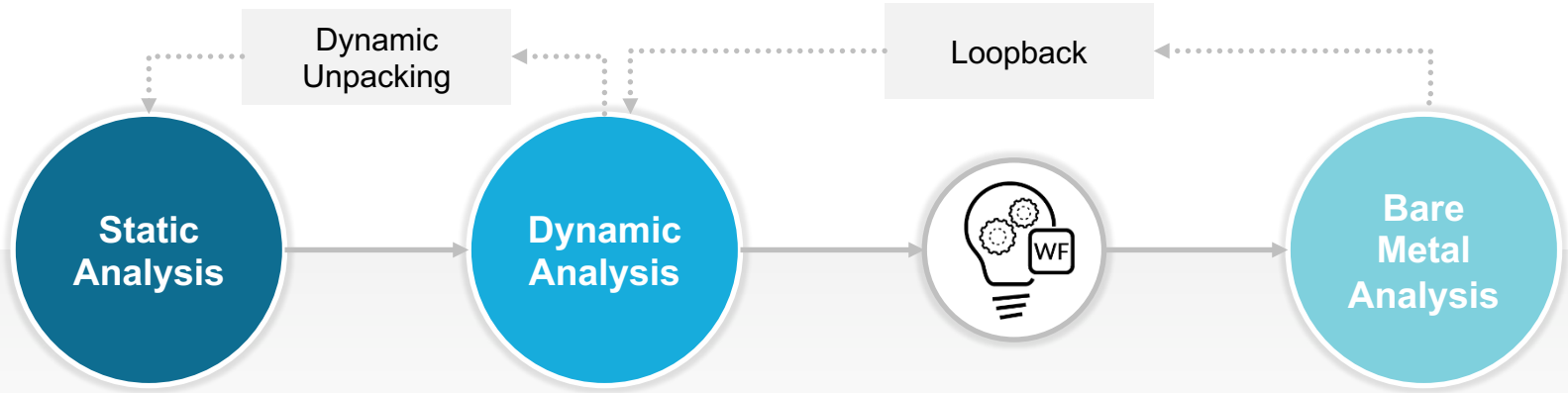
WildFire going beyond sandboxing



Wildfire Cloud Service @ Palo Alto Networks

- *Dynamic analysis*
- *Static analysis*
- *Machine learning*
- *Bare metal analysis*

Wildfire Cloud Service @ Palo Alto Networks



Detect known exploits, malware, and variants

- Memory analysis
- Machine learning
- File anomalies
- Malicious patterns
- Known malicious code

Find new zero-day exploits & malware through execution

- Network Traffic Profiling
- Custom hypervisor
- Behavioral scoring
- Multi-version analysis

Heuristic Engine
Steer evasive malware to bare metal

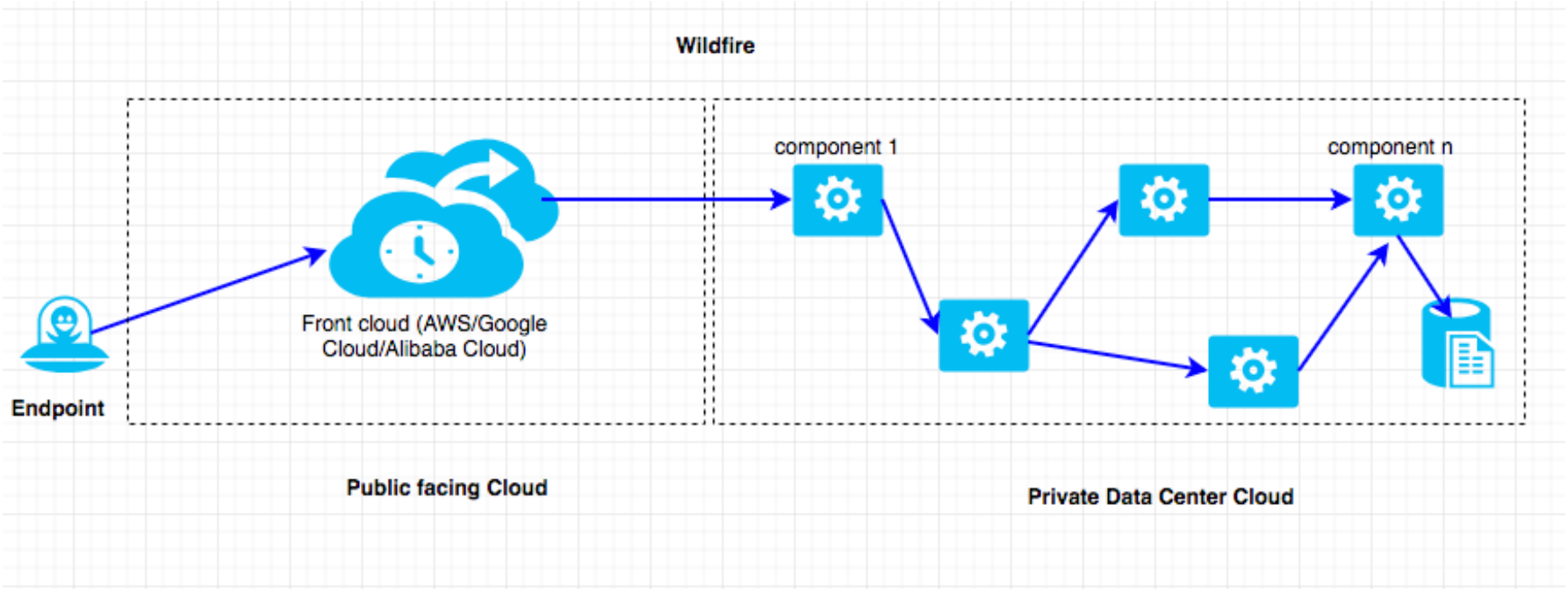
Identify VM-aware threats using hardware systems

- Full dynamic analysis
- Real desktop hardware
- No virtual environment
- No hypervisor

Wildfire Cloud Service @ Palo Alto networks

- *Hybrid cloud*
- *Distributed computation system*
- *Multiple process pipelines*
- *Cross region data centers*
- *Thousands of machines running in the cloud*
- *Serves millions of Palo Alto networks endpoints client daily*

Wildfire Cloud Service @ Palo Alto networks



Pain Points of DevOps

Pain Points of DevOps

- No visibility into production system
- Hard to trace processing entity through its whole life cycle in cloud processing in real time
- Difficult to get accurate real-time processing statistic data for whole cloud system matrix performance
- No ability to track failure components in real time based on processing logic
- Painful to identify the root cause for failure processing case
- Incapable of predicting and preventing from disasters from occurring

Ranger

Streaming based tracking & monitoring system for cloud service

Ranger- Goals

- Provide high visibility to production system
- Real-time to track and monitor processing pipeline in Cloud
- Trace process entity in its whole life cycle in Cloud
- Identify the failure case with context information
- Narrow down the malfunctional components based on application logic in fast pace
- Provide accurate statistic data for over all cloud system and individual components
- Support high throughput
- Build up production data for forensic analysis

Ranger-Solutions

- Streaming technology is the way to go to achieve these goals
  - Twitter storm

    Storm is first generation streaming system, and it was launched and open sourced by twitter. Not very popular now.
  - Spark streaming system

    Near real time system, recently support continuous streaming which is close to real time stream, but not available for production quality.
  - Apache beam

    Apache beam is established and created by google, it is an abstract layer on the streaming process engine. It is well designed and suitable for running on different streaming process engine like google data flow, storm and flink. It still has some gap to support 3rd party streaming process engine but more friendly for google data flow engine by nature.
  - Flink

    Flink is most popular and mature streaming process engine in the world now. It has couple highlights like high performance, support exactly once semantic, support process time and event time model.
  - Kafka Streams

    Pretty new and promising, and establish as eco-system including data pop in and data sink plus streaming handle

Ranger-Solutions

- Kafka Streams 
  - Many companies adopted Kafka as its major queue system
  - Proven high throughput, scalable and HA system
  - Streaming feature is new but evolved fast
  - Powerful and support exactly-once semantics
  - Very active community
  - Make whole infrastructure **simple** and maintenance easy
  - Easily to hook up other systems through Kafka Connector
  - Existing components in Cloud system at Palo Alto Networks

Ranger-Solutions

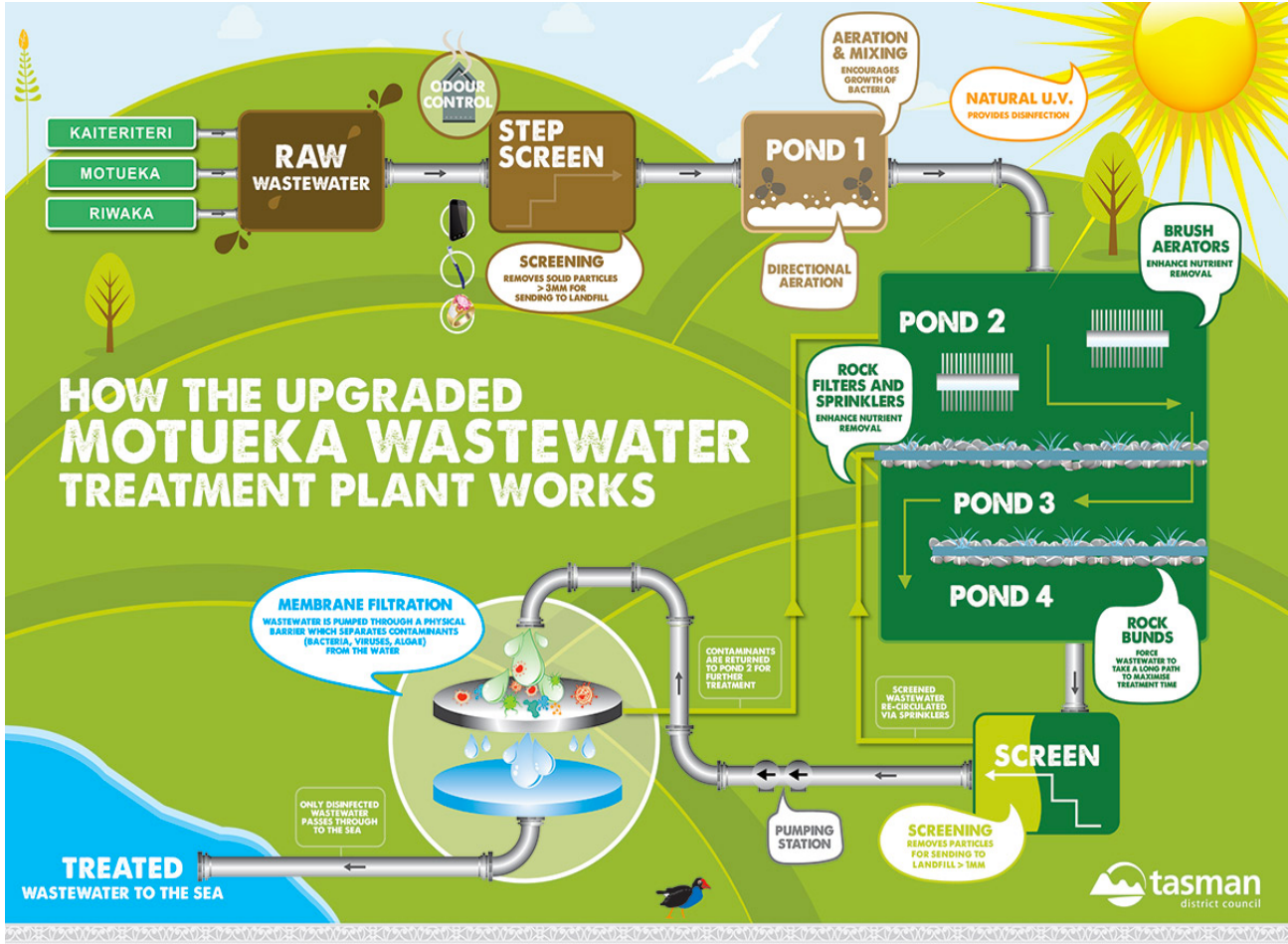
- Ranger @ Palo Alto Networks

Ranger-Solutions

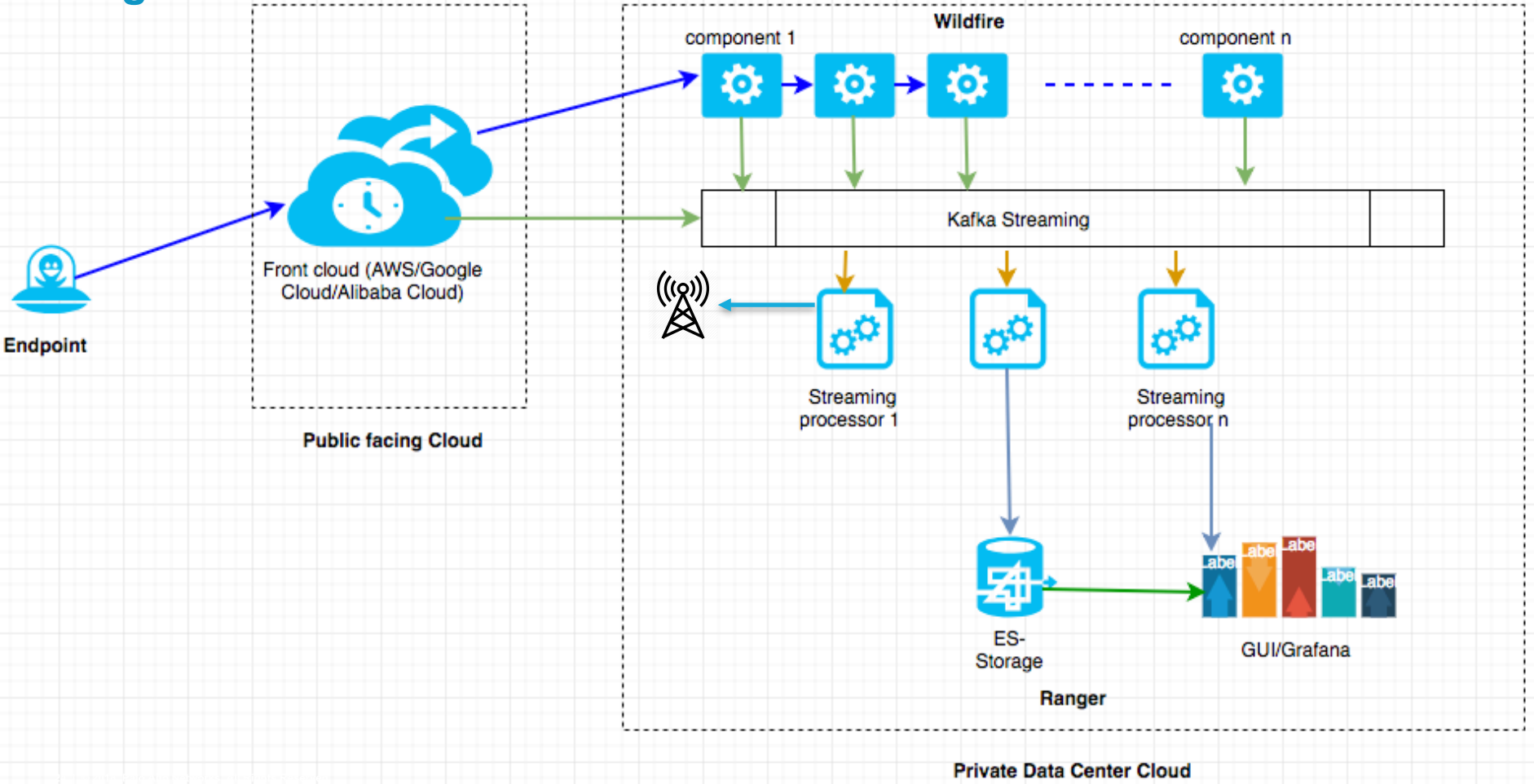
- Kafka as streaming backbone
- Home grown framework to collect and generate ranger log which is used to trace and monitor cloud services
- Kafka streaming process engine to parse, analyze and aggregated the ranger log
- Elasticsearch to store sanitized raw data, statistic data and other result data
- Customized dashboard GUI to display real time result
- Hook up with Grafana to present historical statistic result
- Open-tracing standard compatible (scoped in next release)
- Prometheus to gather system level matrix data (CPU usage, networks and memory usage etc, scoped in next release)
- Instrument (Filebeat) ranger log from every components into ranger streaming engine

Ranger-Solutions

- How ranger solution work

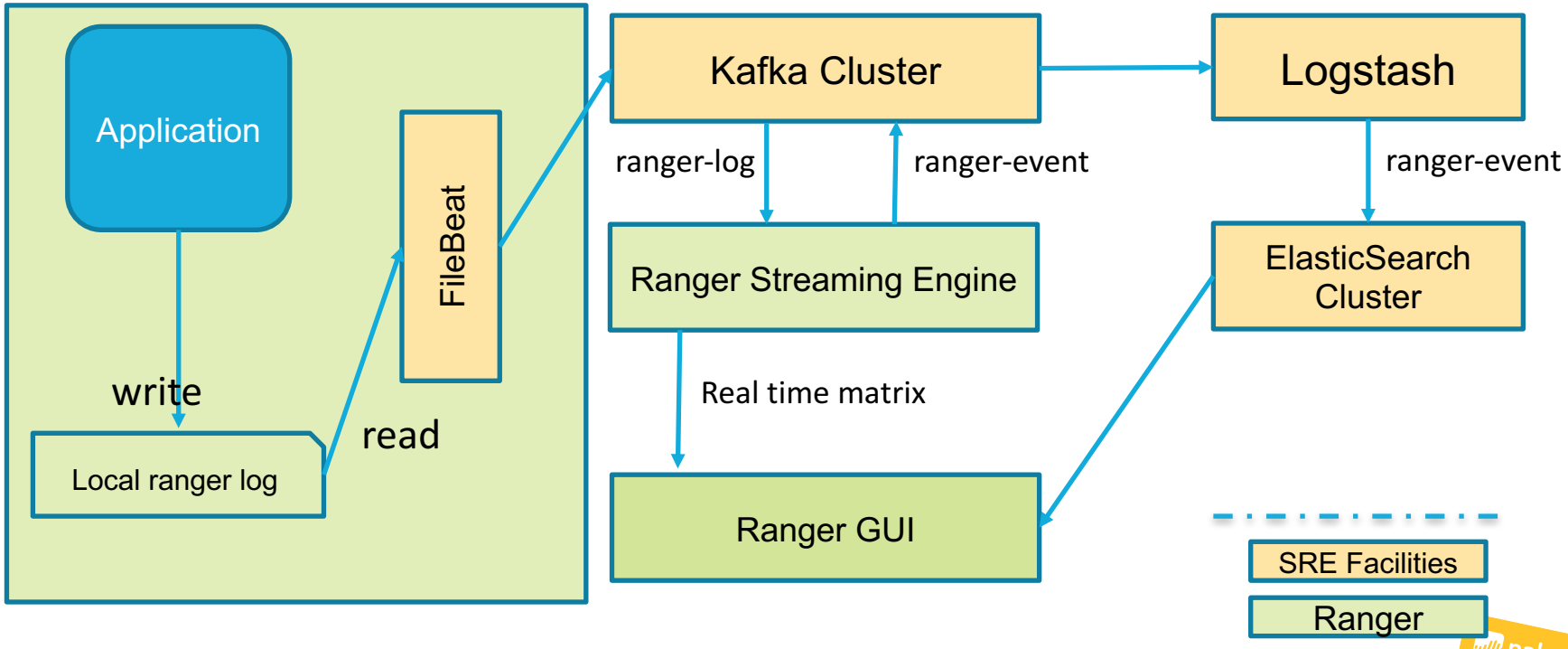


Ranger-Solutions



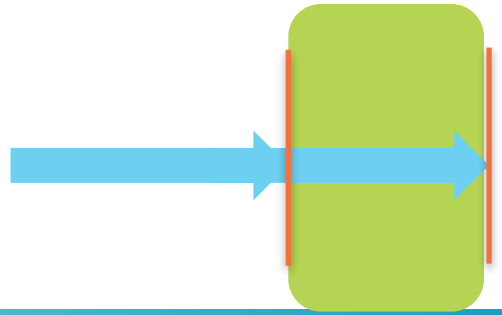
Ranger-Solutions

32 Categories of WF components



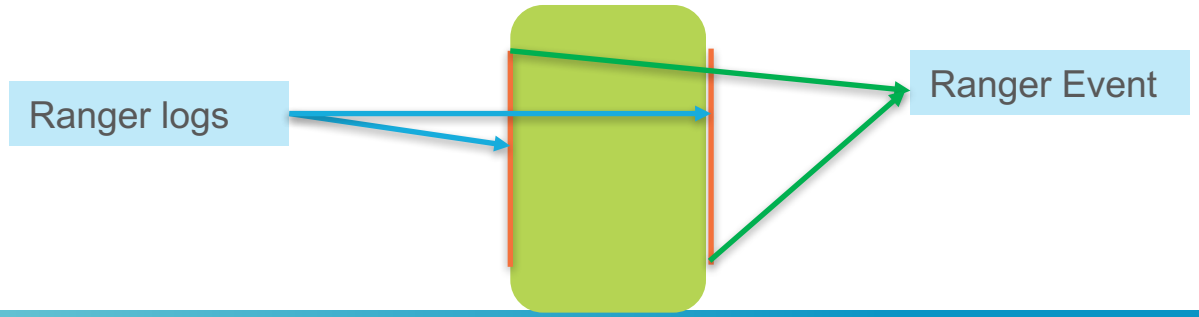
Ranger-Solutions

- Generate unique trace id when request enters cloud
- Data flow in one component, mark a footprint as a **ranger log**
- Data flow out of the component, mark a footprint as a **ranger log**
- Two footprint composes of **one ranger event** flowing through the component
- Ranger event contains many meta information like the component process duration, hostname, component type, necessary context information etc.



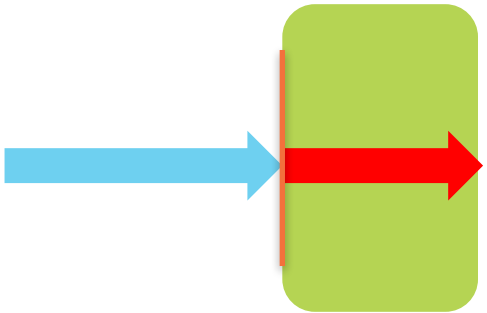
Ranger-Solutions

- **Two pairing Ranger Logs = Ranger Event**



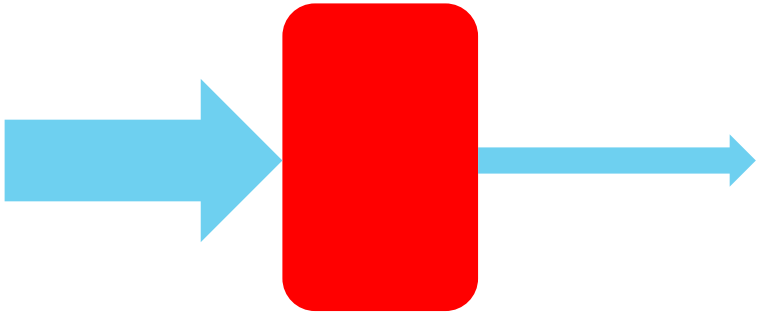
Ranger-Solutions

- Anomaly detection
  - Only have flow in ranger log, no flow out ranger log/or flow out with error flag ranger log within reason timeframe
  - Mark the flow in entity as anomaly



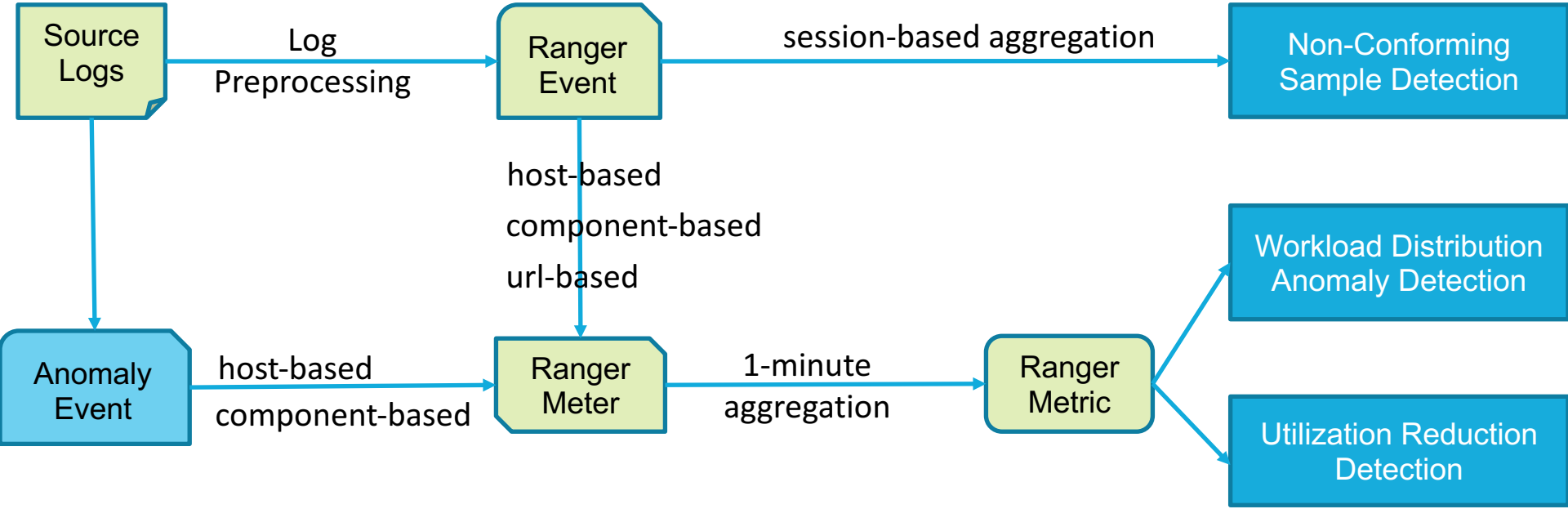
Ranger-Solutions

- Component anomaly detection
  - Anomaly events increase dramatically within threshold time period
  - Anomaly events has trend to continue increasing



Ranger-Solutions

- Streaming-Based Anomaly Detection

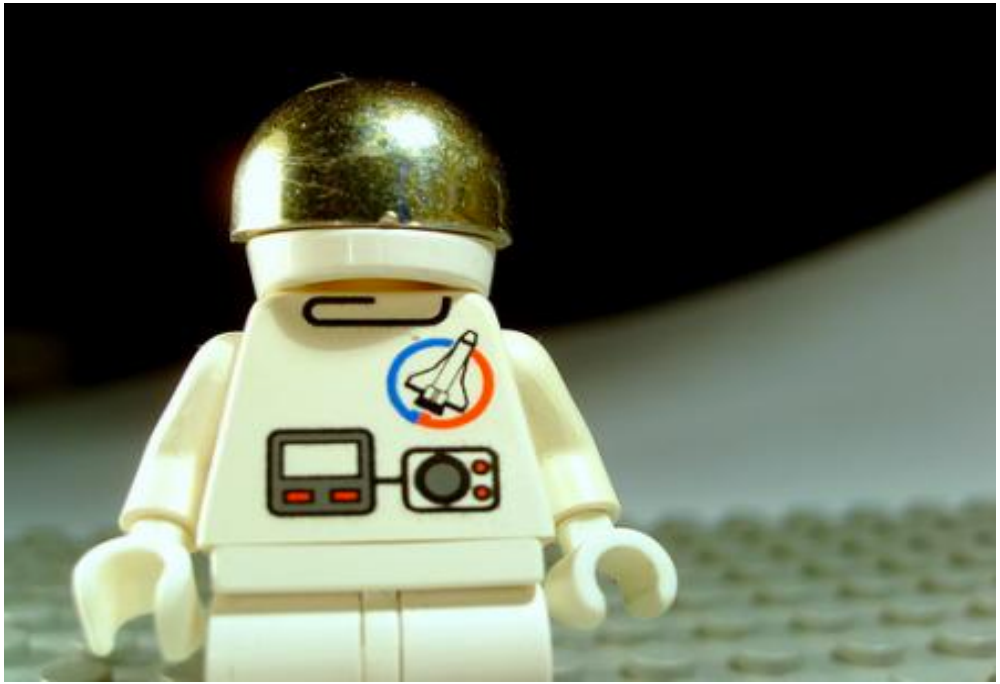


Ranger-Result

- Already in production
- Be able to handle up to 300M/hour throughput traffic
- Linearly scale out Ranger streaming process engine
- Increase the whole cloud services visibility
- Helped identify some critical application defect logic in the first week deployment
- Speeded up customer issues resolve by quickly identifying the issue root cause
- Improved overall cloud service visibility and quality

Ranger-Result

A small step, change world



Ranger-Result- Case study Old way

Ranger-Result- Today

Ranger-Result- Case study

From passive action to proactive prevention

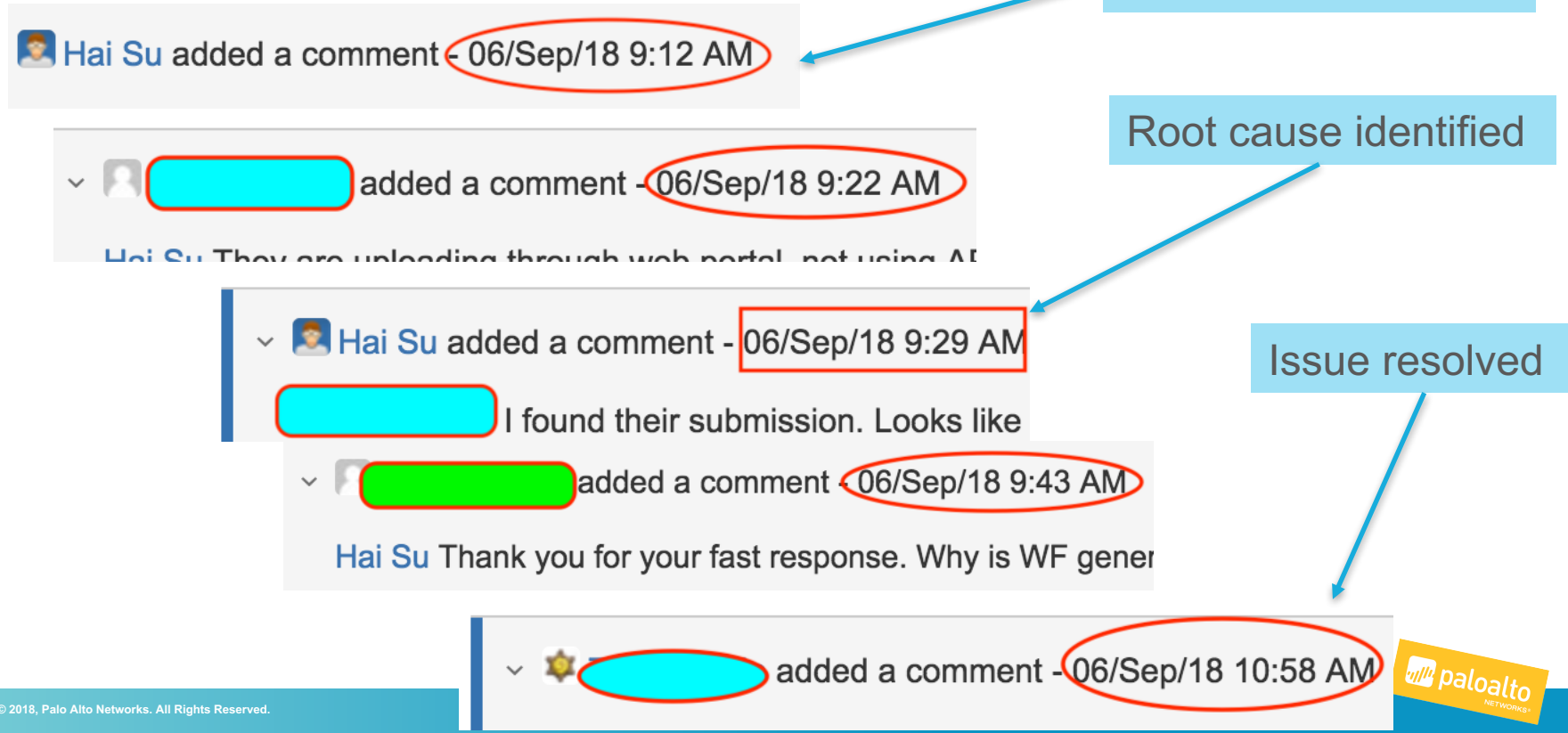
- Suddenly see spike of anomaly processing every minute
- Deep dig out most of processing landing on one service instance
- Service log revealing error on data processing
- Isolate the instance from receiving more traffic
- Escalate to right team to investigate
- Finding out that instance has wrong package installed

Ranger-Result- Case study

Details

Type: Bug

Priority: P2



Ranger-Result--Summary

- **Ranger data provenance**
  - ✓ Provides visibility into cloud operation environment
    - historical view of WF samples, their origin and processing activities
    - the distribution of workloads across thousands of components and servers.
  - ✓ Enables analytics capabilities to detect system anomaly
    - Elasticsearch serves as a data warehouse like source for batch processing
    - Kafka Streams enables near real time streaming processing

- **Extensible Data Collecting Mechanism**
  - ✓ Provides rich application-level information
    - URL-based statistics
    - Sample type statistics
  - ✓ Enables the possibility to build new application
    - Near real-time sample access to train machine learning algorithms
    - Statistical report of to evaluate analyzer performance - Static vs. dynamic detection rates and analytics

Future work

Future work

- Ranger build up a Gold Mine with production application data
- Make everything possible to improve SRE work life balance
  - Utilize Machine learning/AI and container technologies to achieve
  - Auto-Scaling
    - Allocate more resource to potential bottleneck components
  - Auto-Healing
    - Isolate the malfunctional components
  - Resource allocation optimization
    - Put more resources to more demanding components

Future work-isolate mal functional instance

- Stop data flow into the struggling component instance
- Using container technology to graceful shutdown the mal-functional component instance
- Hook up pager system to allow human to be involved

Future work

Q/A

极客时间VIP年卡

每天6元，365天畅看全部技术实战课程

- 20余类硬技能，培养多岗多能的混合型人才
- 全方位拆解业务实战案例，快速提升开发效率
- 碎片化时间学习，不占用大量工作、培训时间

THANK YOU

hai.su@paloaltonetworks.com