Intelligent Cloud Operations
Part 1. Introduction

Definition (Gartner) [**AIOps**]

AIOps platforms utilize big data, modern machine learning and other advanced analytics technologies to directly and indirectly enhance IT operations (monitoring, automation and service desk) functions with proactive, personal and dynamic insight.

Prof. Jorge Cardoso
E-mail: jorge.cardoso@huawei.com
Intelligent Cloud Operations/SRE Dept.
Munich and Ireland Research Centers

2020

Intelligent Cloud Operations
The emerging field of AIOps

The field of **AIOps**, also known as **Artificial Intelligence for IT Operations**, uses advanced technologies to dramatically improve the monitoring, operation, and troubleshooting of distributed systems. Its main premise is that operations can be automated using monitoring data to reduce the workload of operators (e.g., SREs or production engineers). Our current research explores how AIOps – and many related fields such as **deep learning, machine learning, distributed traces, graph analysis, time-series analysis, sequence analysis, advanced statistics, NLP and log analysis** – can be explored to effectively **detect**, **localize**, **predict**, and **remediate failures** in **large-scale cloud infrastructures** (>50 regions and AZs) by analyzing **service management data** (e.g., distributed traces, logs, events, alerts, metrics). In particular, this talk will describe how a particular monitoring data structure, called distributed traces, can be analyzed using deep learning to identify anomalies in its spans. This capability empowers operators to quickly identify which components of a distributed system are faulty.

- Planet/large-scale Distributed systems and cloud computing
- Distributed traces, logs, events, alerts, metrics, ….
- Big Data platforms with Kubernetes, Hadoop, Spark, Flink, …
- Deep Learning, machine learning, data mining, advanced statistics, time-series analysis, NLP, ….
- Detect, localize, predict, and remediate failures in infrastructures

Dr. Jorge Cardoso is Chief Architect for Planet-scale AIOps at Huawei's Ireland and Munich Research Centers. Previously he worked for several major companies such as SAP Research (Germany) on the Internet of Services and the Boeing Company in Seattle (USA) on Enterprise Application Integration. He previously gave lectures at the Karlsruhe Institute of Technology (Germany), University of Georgia (USA), University of Coimbra and University of Madeira (Portugal). His current research involves the development of the next generation of AIOps platforms, Cloud Operations and Analytics tools driven by AI, Cloud Reliability and Resilience, and High Performance Business Process Management systems. He has a Ph.D. in Computer Science from the University of Georgia (USA).

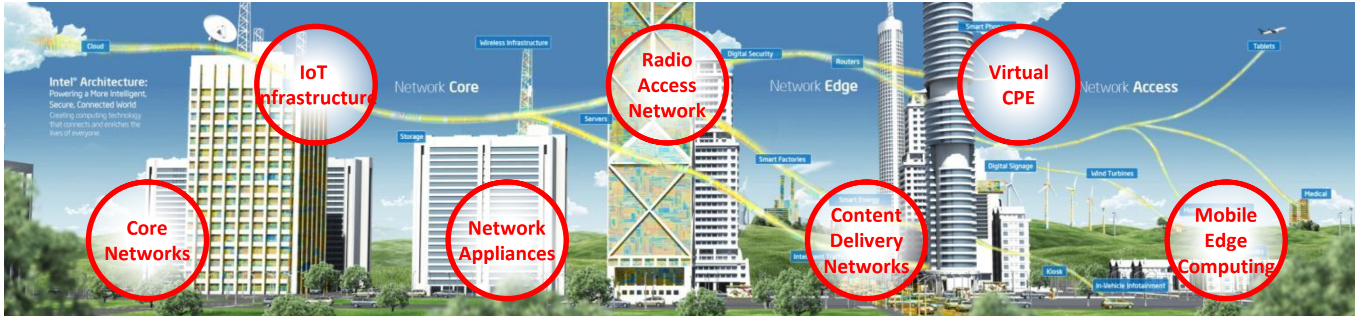
Interests: **AIOps, Service Reliability Engineering, Cloud Computing, Distributed Systems, Business Process Management**

GitHub | Slideshare.net | GoogleScholar

Operations and Maintenance
Global Trends

The cloud has evolved from a disruptor to an obvious solution for enterprise IT. However, the increasing complexity of public, private, edge, and hybrid cloud environments presents new challenges for cost-efficient operations and maintenance (O&M) tasks

Networking



Energy

East Environment Energy

Manufacturing

Geely

Healthcare

PHILIPS Medical

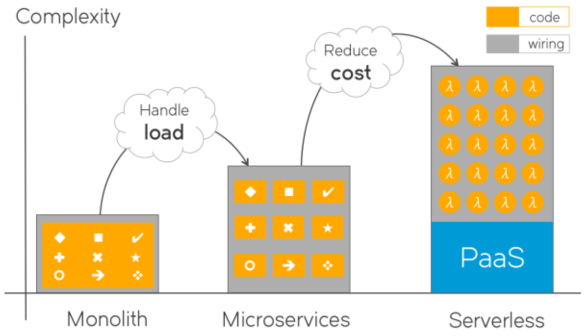
Smart Cities/Buildings

Underlying Problem
Increasing Complexity of Distributed Systems

Every year the management of IT O&M is more complex

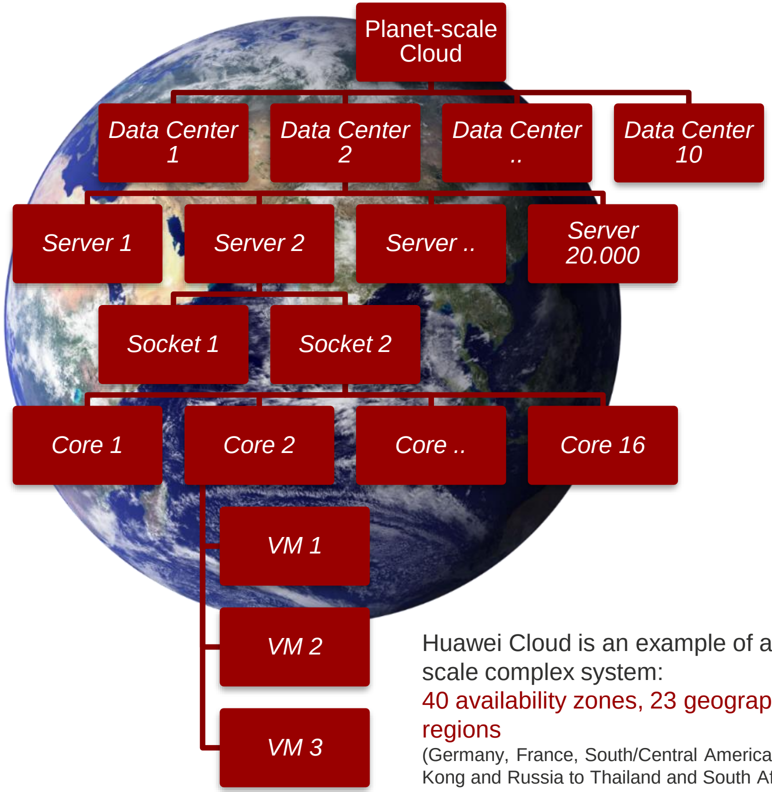
- **Digitalization with cloud, mobile, and edge**
- **Large-scale microservice and serverless systems**
- **Increase in IT size, and event/alert volumes**
- **SLA guarantee for critical applications**
- **Nano/microservices break existing monitoring tools**

However, existing O&M tools still rely on old approaches which involve manual rule configuration, averages and std, and data wrangling to manage physical assets



Microservices/serverless add complexity to communications between services making systems far more difficult to maintain and troubleshoot. With a monolith application, when there is a problem, only one product owner needs to be contacted.

- **An hypothetical planet-scale Cloud**



Huawei Cloud is an example of a planet-scale complex system:
40 availability zones, 23 geographical regions
(Germany, France, South/Central America, Hong Kong and Russia to Thailand and South Africa)

20M VMs (3 VM/core), billion metrics
(for illustrative purpose only)

Intelligent O&M
Commercial Solutions

< 2000 — Centralized Processing

Phase 1.0

- Small data: MB/GB
- Manual operation with Unix tools (ping, traceroute, tcpdump)
- Centralized monitoring tools
- Individual monitoring of servers

2010 — Distributed Processing

Phase 2.0

- Big Data processing: TB/PB
- Integration of distributed processing platforms
- Monitoring of SOA
- Visualization

2012 — Services all-in-one

Phase 3.0

- Automation
- Monitoring as a service
- Cloud / on premise
- Servers, services
- Hadoop, analytics

2015 — AI-based Platforms

Phase 4.0

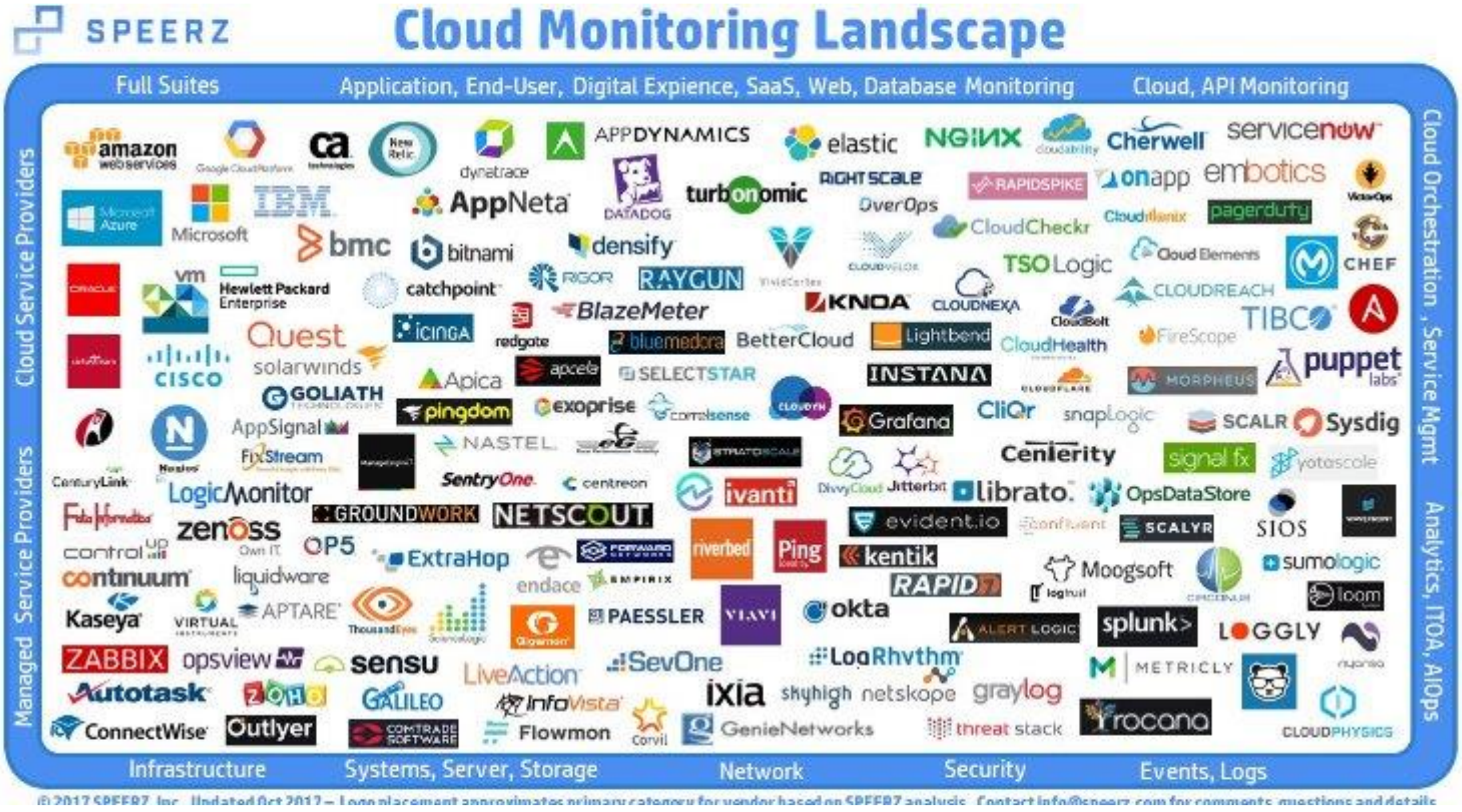
- Enhance existing solutions with AI
- Simple AI algorithms clustering, regressions, outlier detection

2020 — AIOps Platforms

Phase 5.0

- Make AI actionable
- Exploit knowledge graphs
- Fault prediction
- Recommendations for fault handling
- Automated recovery
- Self-service
- Customer service
- Deployment analysis
- Target availability, efficiency, and cost

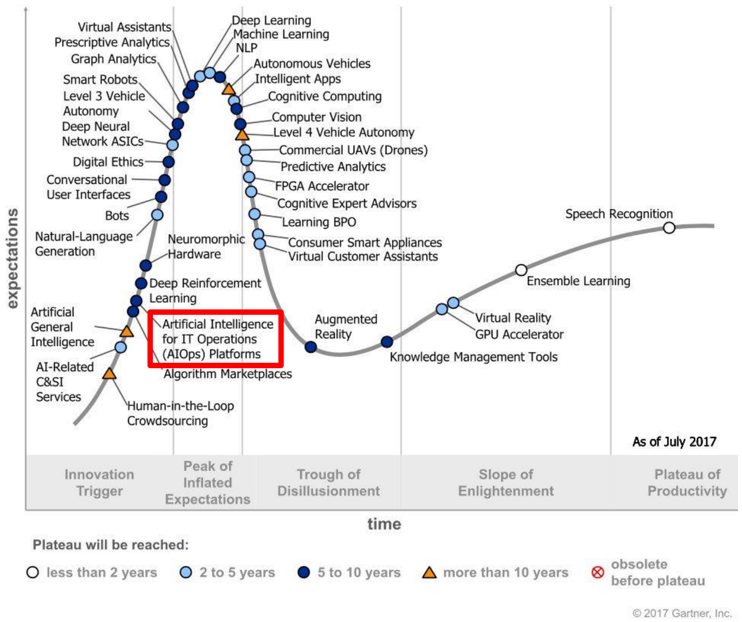
Intelligent O&M
Commercial Solutions



Industry Trends
Bringing AI to O&M

O&M Activities
- **System monitoring** and 24x7 technical support, Tier 1-3 support
- **Troubleshooting and resolution of operational issues**
- Backup, restoration, archival services
- Update, distribution and release of software
- Change, capacity, and configuration management
- …

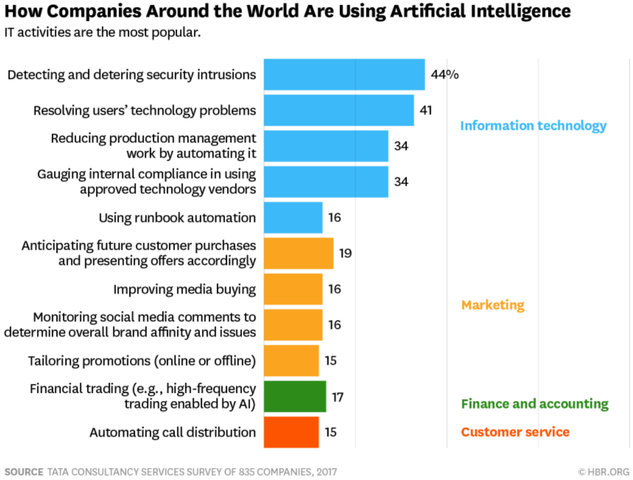
Figure 1. Hype Cycle for Artificial Intelligence, 2017



"We began applying machine learning two years ago **(2016)** to **operate our data centers more efficiently**… over the past few months, DeepMind researchers began working with Google's data center team to significantly improve the system's utility. Using neural networks trained on different operating scenarios and parameters, we created a more efficient and adaptive framework to understand data center dynamics and optimize efficiency." *Eric Schmidt, Dec. 2018*

"Virtyt Koshi, the EMEA general manager for virtualization vendor Mavenir, reckons Google is able to run all of its data centers in the Asia-Pacific with only about 30 people, and that a telco would typically have about 3,000 employees to manage infrastructure across the same area."

Reducing O&M Costs
Google: 30 operators
Others: 3000 operators



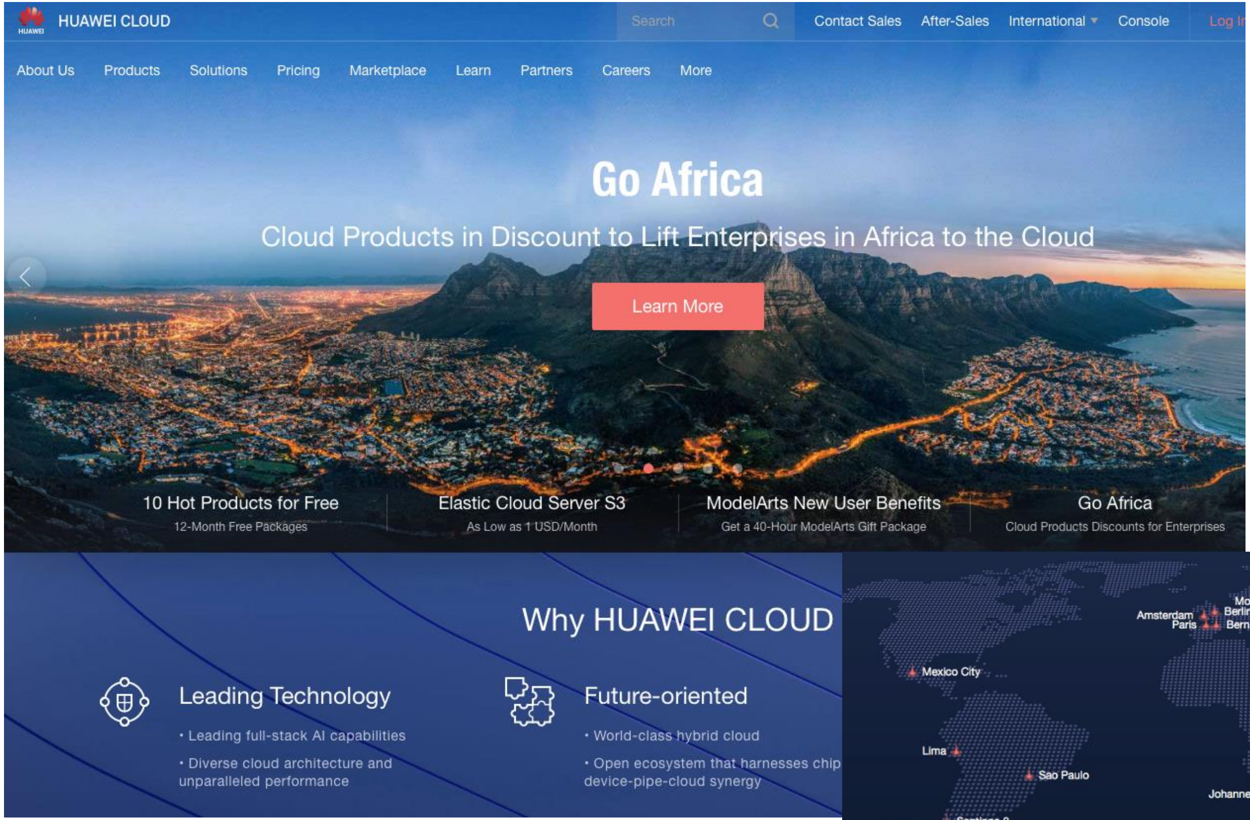
Harvard Business Review

Early indicators
Moogsoft AIOps, Amazon EC2 Predictive Scaling, Azure VM resiliency, Amazon S3 Intelligent Tiering

38.4% of organizations take more than 30 minutes to resolve IT incidents that impact consumer-facing services (PagerDuty)

https://www.lightreading.com/automation/google-has-intent-to-cut-humans-out-of-network/d/d-id/742158
https://www.lightreading.com/automation/automation-is-about-job-cuts---lr-poll/d/d-id/741989
https://www.lightreading.com/automation/the-zero-person-network-operations-center-is-here-(in-finland)/d/d-id/741695

Huawei Cloud
Planet-scale Landscape

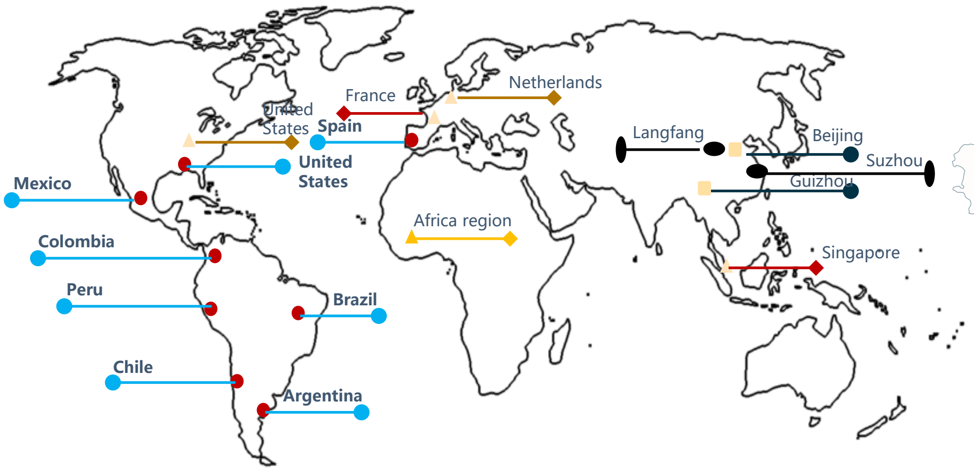


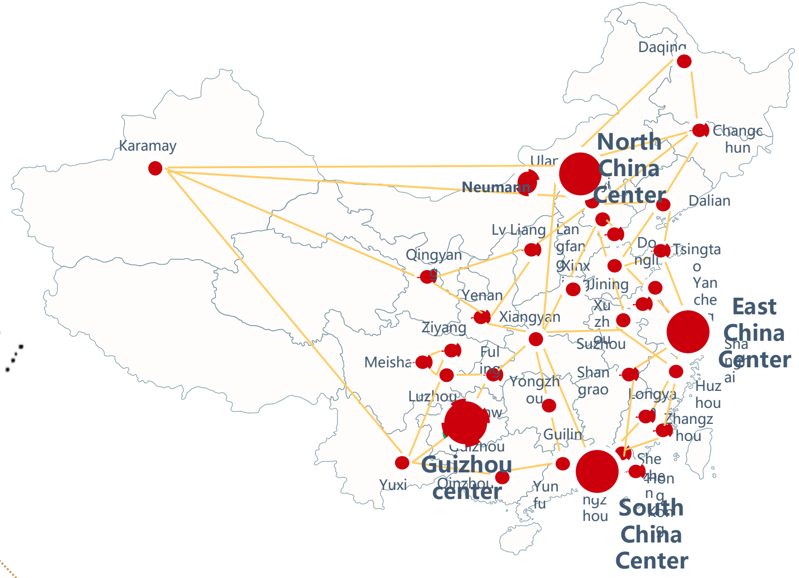
… a few numbers…

- worldwide, Huawei Cloud has 44 availability zones across 23 regions (June 2019)
- more than 180 cloud services and 180 solutions for a wide range of sectors
- Costumers include European Organization for Nuclear Research (CERN), PSA Group, Shenzhen Airport, Port of Tianjin, …

Huawei Cloud
Worldwide Coverage

A global cloud network with resources and service teams all over the world





2015.07 2016.02 2016.06 2016.09 2017.03

- Overseas public cloud + domestic joint operation cloud + domestic public cloud
- Compliance with local laws and regulations, facilitating global expansion of enterprise business
- Three regional centers in China and more than 30 city nodes

Huawei Cloud Success Stories

Huawei Desktop Cloud Helps Hong Kong Airlines Flies High

HKA purchased Huawei's proven cloud-desktop technologies to reduce operating costs, improve work...
Learn More

Serverius Data Centers Get Huawei AntiDDoS

Huawei enables Netherlands Serverius data centers to protect customers against DDoS attacks.
Learn More

New Large Data Center in Sweden with Decreased Environmental Footprint

Huawei helps Binero raise its infrastructure reliability and lower its environmental footprint by...
Learn More

Qatar Airport Adopts Secure Cloud Storage

OceanStor 9000-based video cloud solution securely managed video for Hamad International Airport.
Learn More

Huawei CloudCampus Helps Honda Agency Grow

Huawei CloudCampus Solution enables Dongfeng Honda to lower cost of network, expand dealerships
Learn More

Huawei Helps COFCO Coca-Cola Build an Enterprise Private Cloud Platform

You probably didn't know that, in China, when you drink Coca-Cola, produced by COFCO Coca-Cola...
Learn More

Huawei Cloud Streamlines Beijing Services

Huawei's Distributed Cloud Data Center helps Beijing government speed services for citizens
Learn More

NHS in the UK Constructs an End-to-End Private Cloud Data Center

Huawei helps Avon and Wiltshire Mental Health Partnership NHS Trust (AWP) build an End-to-End (E2E)...
Learn More

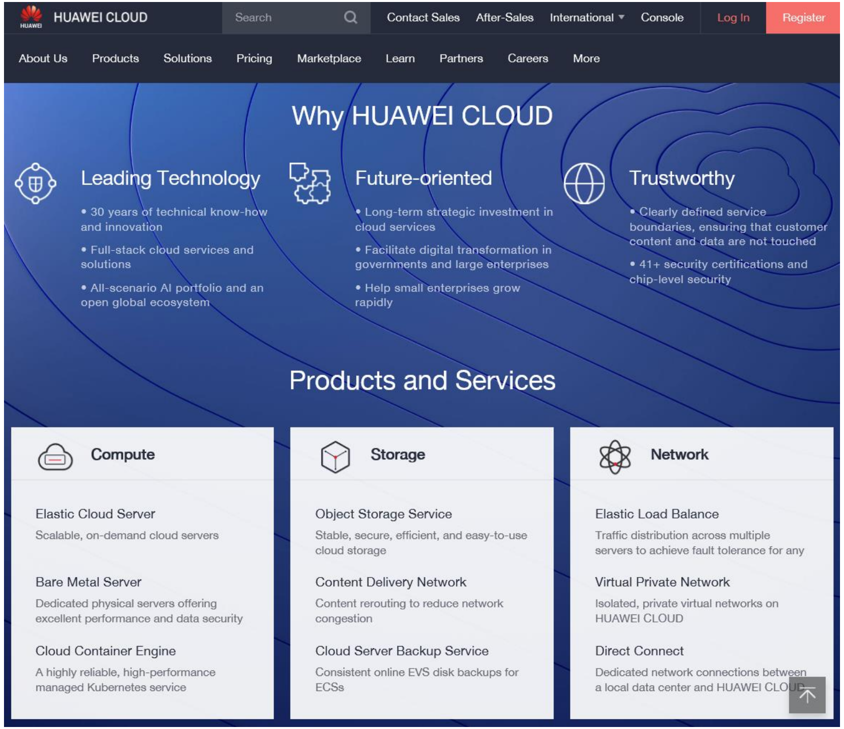
TF1 Finds Performance in Secure Cloud Media

With Huawei's media cloud, TF1 gets secure, high-performance video editing on low-cost terminals.
Learn More

Huawei Data Center Network Solution Assists Ikoula for Cloud Hosting Services

Huawei Data Center Network Solution reduces service deployment complexity, improves system reliabili...
Learn More

AIOps @ Huawei Cloud
Troubleshooting Scenarios



https://intl.huaweicloud.com/

HC System's Components

- OBS. Object Storage Service is a stable, secure, efficient, cloud storage service
- EVS. Elastic Volume Service offers scalable block storage for servers
- VPC. Virtual Private Cloud enables to create private, isolated virtual networks
- ECS. Elastic Cloud Server is a cloud server that provides scalable, on-demand computing resources

Troubleshooting Scenarios

- **Response Time Analysis**
  - A service started responding to requests more slowly than normal
  - The change happened suddenly as a consequence of regression in the latest deployment
- **System Load**
  - The demand placed on the system, e.g., REST API requests per second, increase since yesterday
- **Error Analysis**
  - The rate of requests that fail -- either explicitly (HTTP 5xx) or implicitly (HTTP 2xx with wrong content) -- is increasing slowly, but steadily
- **System Saturation**
  - The resources (e.g., memory, I/O, disk) used by key controller services is rapidly reaching threshold levels

AIOps @ Huawei Cloud Troubleshooting Tasks

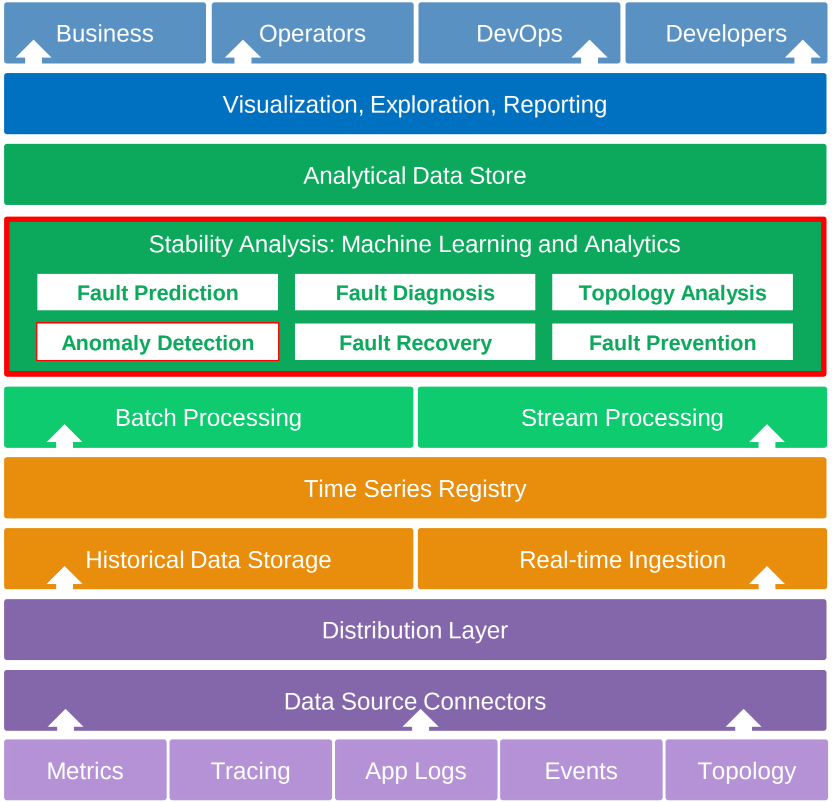
iForesight is an intelligent new tool aimed at SRE cloud maintenance teams. It enables them to quickly detect anomalies thanks to the use of artificial intelligence when cloud services are slow or unresponsive.

O&M Troubleshooting Tasks

Anomaly Detection	Fault Diagnosis	Fault Prediction	Fault Recovery

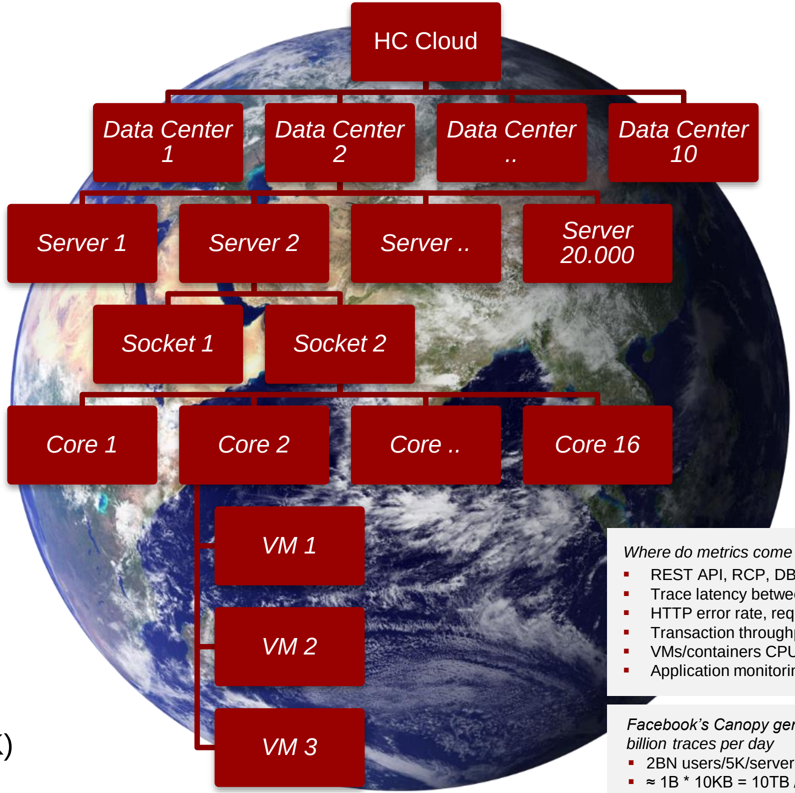
- **Anomaly detection**. Determine what constitutes normal system behavior, and then to discern departures from that normal system behavior
- **Fault diagnosis (root cause analysis)**. Identify links of dependency that represent causal relationships to discover the true source of an anomaly
- **Fault Prediction**. Use of historical or streaming data to predict incidents with varying degrees of probability
- **Fault recovery (Incident Remediation)**. Explore how decision support systems can manage and select recovery processes to repair failed systems

Reference architecture for AIOps platforms



Challenge: Planet Scale Monitoring Data Generated

- **An hypothetical planet-scale Cloud**
  - ** 10 data centers (~10MW of power)*
  - ** 20.000 server (500 watts)*
  - ** 2 sockets * 16 cores*
  - ** 3 VMs*
  - *= 200K servers*
  - *≈ 20M VMs*
- **Infrastructure and Services**
  - 200K servers
  - * 75 (CPU, MEM, disk, net, processes, etc.)
  - + 25 (counters, gauges, trace, log, metrics)
  - = 200K * 100 = 20M time-series
- **VMs**
  - 20M VMs
  - * 75 (CPU, MEM, disk, net, processes, etc.)
  - + 25 (counters, gauges, trace, log, metrics)
  - ≈ 20M * 100 = 2B time-series
- **Traffic**
  - > 100 bytes per observation
  - ≈ 20M * 100 = 2GB / sec ≈ 200TB / day (≈ *100K)
  - ≈ 2B * 100 = 200GB / sec ≈ 2PB / day (≈ *100K)



Where do metrics come from

- REST API, RCP, DB calls latency
- Trace latency between events
- HTTP error rate, requests
- Transaction throughput
- VMs/containers CPU, MEM, DISK
- Application monitoring data

Facebook's Canopy generates and processes 1.3 billion traces per day

- 2BN users/5K/server = 360K servers
- ≈ 1B * 10KB = 10TB / day

AIOps @ Huawei Cloud
Monitoring Data Structures

System's Components (e.g., compute, storage, network) are monitored and generate various types of data: Logs, Metrics, Traces, Events, Topologies

Logs. Service, microservices, and applications generate logs, composed of timestamped records with a structure and free-form text, which are stored in system files.

2017-01-18 15:54:00.467 32552 ERROR oslo_messaging.rpc.server [req-c0b38ace - default default] Exception during message handling

Metrics. Examples of metrics include CPU load, memory available, and the response time of a HTTP request.

{“tags": [“mem”, “192.196.0.2”, “AZ01”], “data”: [2483, 2669, 2576, 2560, 2549, 2506, 2480, 2565, 3140, …, 2542, 2636, 2638, 2538, 2521, 2614, 2514, 2574, 2519]}

Traces. Traces records the workflow and tasks executed in response to an HTTP request.

{"traceId": "72c53", "name": "get", "timestamp": 1529029301238, "id": "df332", "duration": 124957, “annotations": [{"key": "http.status_code", "value": "200"}, {"key": "http.url", "value": "https://v2/e5/servers/detail?limit=200"}, {"key": "protocol", "value": "HTTP"}, "endpoint": {"serviceName": "hss", "ipv4": "126.75.191.253"}]

Events. Major milestones which occur within a data center can be exposed as events. Examples include alarms, service upgrades, and software releases.

{"id": "dns_address_match“, "timestamp": 1529029301238, …}
{"id": "ping_packet_loss“, "timestamp": 152902933452, …}
{"id": "tcp_connection_time“, "timestamp": 15290294516578, …}
{"id": "cpu_usage_average “, "timestamp": 1529023098976, …}

Planet Scale
Monitoring Data Generated

> *" Operating planet-scale, globally distributed cloud platforms requires accurate monitoring of the health and performance of systems which cannot be handled by traditional commercial solutions "*

- ► Facebook uses Gorilla to stores up to **10 billion unique time series**,**1 trillion data points per day**, and servers **18 million queries per minute.**
- ► Twitter monitoring and troubleshooting service handles more than **2.8 billion write requests per minute**, stores **4.5 petabytes of time series data**, and handles **25k query requests per minute**.

Challenges for Huawei Cloud BU

- *State transition*. Identify issues that emerge from a service updates and configuration change that result in a significant state transition
- *High availability*. Even if a network partition disconnects different datacenters, systems within a datacenter should be able to write data to local machines and retrieve data on demand.
- *Writes domination*. Support a write rate greater that tens of millions of data points each second.
- *Fault tolerance*. Replicate all writes to multiple regions to survive the loss of any given datacenter or geographic region due to a disaster.

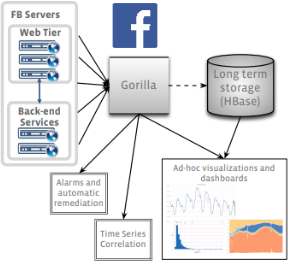
Gorilla: A Fast, Scalable, In-Memory Time Series Database

Tuomas Pelkonen Scott Franklin Justin Teller
Paul Cavallaro Qi Huang Justin Meza Kaushik Veeraraghavan
Facebook, Inc.
Menlo Park, CA

ABSTRACT

Large-scale internet services aim to remain highly available and responsive in the presence of unexpected failures. Providing this service often requires monitoring and analyzing tens of millions of measurements per second across a large number of systems, and one particularly effective solution is to store and query such measurements in a time series database (TSDB).

A key challenge in the design of TSDBs is how to strike the right balance between efficiency, scalability, and reliability. In this paper we introduce Gorilla, Facebook's in-memory TSDB. Our insight is that users of monitoring systems do not place much emphasis on individual data points but rather on aggregate analysis, and recent data points are of much higher value than older points to quickly detect and diagnose the root cause of an ongoing problem. Gorilla optimizes for remaining highly available for writes and reads, even in the face of failures, at the expense of possibly dropping small amounts of data on the write path. To improve query efficiency, we aggressively leverage compression techniques such as delta-of-delta timestamps and XOR'd floating point values to reduce Gorilla's storage footprint by 10x.



How Twitter monitors millions of time series

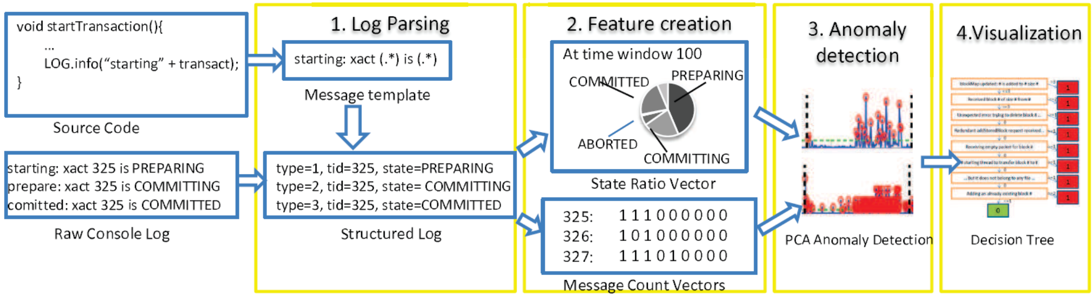
A distributed, near real-time system simplifies the collection, storage, and mining of massive amounts of event data

One of the keys to Twitter's ability to process 500 millions tweets daily is a software development process that values monitoring and measurement. A recent post from the company's *Observability* team detailed the software stack for monitoring the performance characteristics of software services, and alert teams when problems occur. The Observability stack collects 170 million individual *metrics* (time series) every minute and serves up 200 million queries per day. Simple query tools are used to populate charts and dashboards (a typical user monitors about 47 charts).

Application Log Analysis
Error Rate Analysis

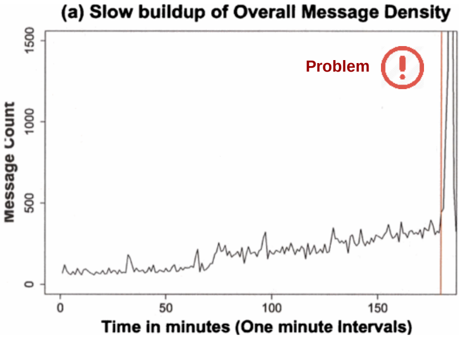
Event Density Anomalies

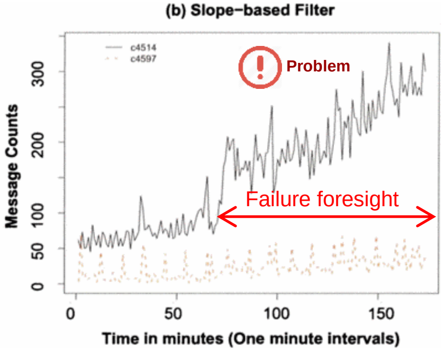
- Integrate operations, running, monitoring, etc. logs
- Similar to latency anomalies but we **replace latency by the number of events generated** over a time window or by a behavioral model
- Use sudden and incremental level shift algorithms
- Use outlier detection algorithms



Approach developed for Google infrastructure [3]

There is a period of time before/during a failure during which the event sources behaves anomalously (see [1]).





Classify the nature of failures

- Permanent fault: remains active for a significant period of time (e.g.: damaged or incorrectly implemented component)
- Transient fault: appears for a very short period of time and disappears (eg: soft error)
- Intermittent (periodic) fault: appears, disappears, and reappears (eg: a parasitic signal emitted by a part of an electronic system disturbs another part during the operation)Transient

[1] A log mining approach to failure analysis of enterprise telephony systems, Dependable Systems and Networks With FTCS and DCC, 2008. DSN 2008. IEEE International Conference on
[2] Google Four Golden Signals for Monitoring Distributed Systems
[3] Wei Xu, Ling Huang, Armando Fox, David Patterson, and Michael I. Jordan. 2009. Detecting large-scale system problems by mining console logs. In Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles (SOSP '09). ACM, New York, NY, USA, 117-132.

Topology Analysis
Exploring Causality for Root Cause Analysis

<u>How to build a topology service/knowledge graph to assist RCA using real-time traces and static CMDB data?</u>

<u>Topology Construction</u>

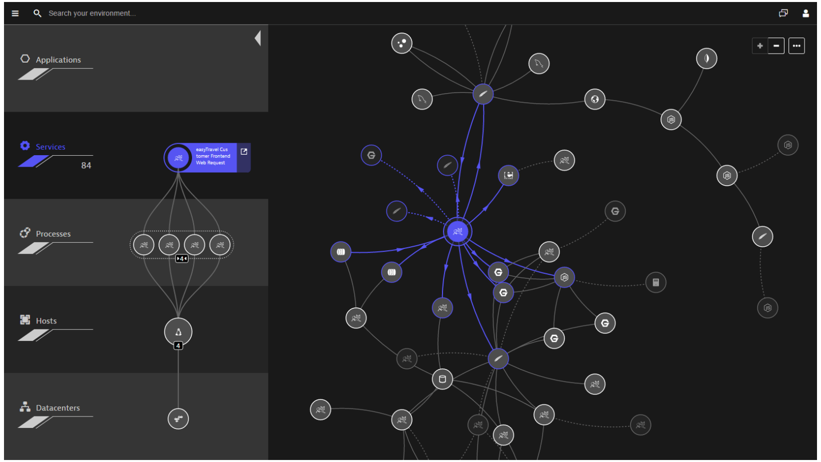
- Discover the dependencies between components (nova, controllers, hosts, services, etc.) using trace information and CMDB
- Construct a **graph** with each node representing a component.
- Since traces have correlation information between two components and timing information, establish **causality**.

<u>Use Cases</u>

- Root Cause Analysis using topologies
- Alarm deduplication using topologies
- Service-to-service communication (Service Networks) analysis
- Trace analysis and reliability

<u>Root Cause Analysis</u>

- Step 1. When a component fails, analyze the log-based behavior model, resource model, and other models of the faulty component that failed.
- Step 2. Search the topology and find all the predecessors of the faulty component.
- Step 3. Analyze the faulty components predecessors. Repeat Step 1.



Dynatrace cloud infrastructure monitoring using topology models using CMDBs

<u>Layers</u>

Operation Layer. Nodes representing the operations which can be invoked

Service Layer. Nodes representing the major services of HC, e.g., ECS and OBS

Microservice Layer. Nodes of microservices supporting major services

Virtual Layer. This layer captures all software-defined entities in the datacenter such as virtual machines (VM), containers, load balancers, virtual networks.

Physical Layer. Physical hosts and associated system resources to which virtual entities are mapped, e.g., memory, disks, NICs, CPUs, etc.

Network Layer. ToR switches, routers, etc.

Infrastructure Layer. Racks, Data Centers, Region, AZ, etc.

Thank you.

Bring digital to every person, home and organization for a fully connected, intelligent world.

Copyright©2019 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

HUAWEI